Estimating the Persistent Homology of $\mathbb{R}^n$-Valued Functions Using Function-Geometric Multifiltrations

Ethan André
École Normale Supérieure, Paris, France

Jingyi Li
Inria Saclay and École Polytechnique, Palaiseau, France

David Loiseaux
Inria Saclay, Palaiseau, France

Steve Oudot
Inria Saclay and École Polytechnique, Palaiseau, France

Abstract

Given an unknown $\mathbb{R}^n$-valued function f on a metric space X, can we approximate the persistent homology of f from a finite sampling of X with known pairwise distances and function values? This question has been answered in the case $n = 1$, assuming f is Lipschitz continuous and X is a sufficiently regular geodesic metric space, and using filtered geometric complexes with fixed scale parameter for the approximation. In this paper we answer the question for arbitrary n, under similar assumptions and using function-geometric multifiltrations. Our analysis offers a different view on these multifiltrations by focusing on their approximation properties rather than on their stability properties. We also leverage the multiparameter setting to provide insight into the influence of the scale parameter, whose choice is central to this type of approach. From a practical standpoint, we show that our approximation results are robust to input noise, and that function-geometric multifiltrations have good statistical convergence properties. We also provide an algorithm to compute our estimators, and we use its implementation to conduct extensive experiments, on both synthetic and real biological data, in order to validate our theoretical results.

2012 ACM Subject Classification Mathematics of computing → Algebraic topology

Keywords and phrases Topological data analysis, multi-parameter persistent homology, function-Rips multifiltration

Digital Object Identifier 10.4230/LIPIcs.SoCG.2026.6

Related Version *Full Version*: https://arxiv.org/abs/2412.04162 [1]

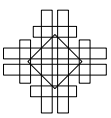
Supplementary Material *Software*: https://doi.org/10.5281/zenodo.19135090

Funding *David Loiseaux*: Inria Action Exploratoire PreMediT (Precision Medicine using Topology).

1 Introduction

Let X be a metric space and $f\colon X \to \mathbb{R}^n$ a function, both unknown. Suppose we are given a finite point cloud P sampled from X, such that the pairwise distances between the points of P are known, as well as the function values at the points of P. Given this input, can we approximate the persistent homology $H_*(f)$ of f (i.e., the persistence module induced in homology from the multifiltration of X by the sublevel sets of f) with provable guarantees? This question was addressed in [10] in the case $n = 1$. The authors proposed an estimator based on a nested pair of Rips complexes $\mathcal{R}^\delta(P) \subseteq \mathcal{R}^{2\delta}(P)$, where δ is a user-defined parameter, which they filtered by the values of f at the vertices using lower-star filtrations. The estimator was then the image of the morphism of 1-parameter persistence modules

© Ethan André, Jingyi Li, David Loiseaux, and Steve Oudot;
licensed under Creative Commons License CC-BY 4.0



induced in homology by the inclusion $\mathcal{R}^\delta(P) \hookrightarrow \mathcal{R}^{2\delta}(P)$. We denote by $H_*(\mathcal{R}^{\delta\to 2\delta}(\hat{f}|_P))$ this estimator. Assuming both the domain X and the function $\hat{f}$ are sufficiently regular (typically, X is a compact geodesic metric space with positive convexity radius ϱ_X, and $\hat{f}$ is c-Lipschitz continuous), and P is an ε-sample of X in the geodesic distance for some small enough value of ε (specifically, $\varepsilon < \varrho_X/4$), they proved that the estimator is $2c\delta$-interleaved with $H_*(\hat{f})$ for any choice of parameter δ within the range $[2\varepsilon, \varrho_X/2)$. Thus, in cases where ε is known or can be estimated reliably, one can get an $O(2c\varepsilon)$-interleaving with the target $H_*(\hat{f})$, hence an $O(2c\varepsilon)$-matching with its barcode, by the algebraic stability theorem [2, 8, 9]. Under a reasonable sampling model, ε goes to zero as the number of sample points goes to infinity, hence so does the approximation error, which means that the proposed estimator is consistent. This approach has since been applied in various contexts, notably in clustering where the algorithm ToMATo [11] is an instance of the method in homology degree 0. However, two fundamental questions were left open in [10, 11]:

Q1 how to estimate ε, or to choose parameter δ directly, when ε is unknown;

Q2 how to generalize the approach to arbitrary n, i.e., to vector-valued functions $\hat{f}$.

Q2 is precisely the question we asked at the beginning. Q1 is related to the general problem of scale parameter estimation, which has received a lot of attention in statistics, including in the context of topological data analysis (TDA) – see e.g. [4, 6, 23].

Contributions

In this paper we address both Q1 and Q2 using a unified persistence-based approach.

For Q2 we study the extension of the estimator $H_*(\mathcal{R}^{\delta\to 2\delta}(\hat{f}|_P))$ from [10] to vector-valued functions $\hat{f}\colon X \to \mathbb{R}^n$, where the codomain $\mathbb{R}^n$ is equipped with the product order, and we generalize its approximation guarantee to this setting. For the sake of generality we assume $\hat{f}$ to be ω-continuous for some modulus of continuity ω, that is: $\|\hat{f}(x) - \hat{f}(y)\|_\infty \leq \omega(d_X(x,y))$ for any $x, y \in X$, where $\omega\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a non-decreasing subadditive function such that $\omega(\delta)$ goes to 0 as δ does. This includes Lipschitz or Hölder continuous functions as special cases. Under this assumption, our first main result (Theorem 3) states that the estimator $H_*(\mathcal{R}^{\delta\to 2\delta}(\hat{f}|_P))$ is $\omega(2\delta)$-interleaved with its target $H_*(\hat{f})$ for any choice of δ within the range $[2\varepsilon, \varrho_X/2)$, under the same regularity condition on X and ε-sampling condition on P as before. Our proof takes inspiration from the one in [10] but uses a new, purely diagrammatic formulation that clarifies it and makes it hold for any $n \geq 1$.

For Q1 we propose two complementary approaches. The first one infers a relevant choice of δ under an appropriate statistical model. This model assumes the points P are sampled i.i.d. according to some unkown probability distribution μ that is supposed to be (a,b)-standard. When parameters a, b are known, we show that there is an explicit optimal choice δ_k for parameter δ, as a function of the sample size k, that makes the sequence of approximations $\left(H_*(\mathcal{R}^{\delta_k\to 2\delta_k}(\hat{f}|_P))\right)_{k\in\mathbb{N}}$ a quasi-minimax estimator of $H_*(\hat{f})$. When a, b are unknown, as would generally be the case in practice, we show that δ_k can be effectively estimated by some $\hat{\delta}_k$ under our statistical model, so that the sequence $\left(H_*(\mathcal{R}^{\hat{\delta}_k\to 2\hat{\delta}_k}(\hat{f}|_P))\right)_{k\in\mathbb{N}}$ is again a quasi-minimax estimator of $H_*(\hat{f})$. These statements are summarized in Theorem 5.

Our second approach to addressing Q1 is meant for cases where our statistical model does not hold and ε cannot be estimated reliably. We follow the general philosophy underpinning persistent homology by letting δ vary over $\mathbb{R}_{\geq 0}$ and treating it as an extra parameter for persistence. This means that our estimator now becomes the $(n+1)$-parameter persistence module $H_*(\mathcal{R}^{\bullet\to 2\bullet}(\hat{f}|_P))$ indexed over $\mathbb{R}_{\geq 0} \times \mathbb{R}^n$, where the notation $\bullet$ indicates that the Rips parameter is also a filtration parameter. In order to enable comparisons, the target $H_*(\hat{f})$

must also be extended to $\mathbb{R}_{\geq 0} \times \mathbb{R}^n$, which we do by taking its left Kan extension $\mathrm{Lan}_\iota H_*(\mathscr{f})$ along the poset embedding $\iota : \mathbb{R}^n \xrightarrow{\cong} \{0\} \times \mathbb{R}^n \hookrightarrow \mathbb{R}_{\geq 0} \times \mathbb{R}^n$. The extension $\mathrm{Lan}_\iota H_*(\mathscr{f})$ merely contains copies of $H_*(\mathscr{f})$ in each vertical hyperplane $\{\delta\} \times \mathbb{R}^n$, connected by horizontal identities, as illustrated in Figure 1. Our third main result (Theorem 6) states that the n-dimensional interleavings happening between $H_*(\mathscr{f})$ and $H_*(\mathcal{R}^{\delta \to 2\delta}(\mathscr{f}|_P))$ inside each vertical hyperplane $\{\delta\} \times \mathbb{R}^n$ for $\delta \in [2\varepsilon, \varrho_X/2)$ commute with the horizontal morphisms in $H_*(\mathcal{R}^{\bullet \to 2\bullet}(\mathscr{f}|_P))$ and in $\mathrm{Lan}_\iota H_*(\mathscr{f})$, so that, within any vertical slab $[2\varepsilon, \delta_0] \times \mathbb{R}^n$ with $\delta_0 < \varrho_X/2$, they all together form a *vertical* $\omega(2\delta_0)$*-interleaving* (i.e., an $\omega(2\delta_0)$-interleaving along the direction $(0, 1, \cdots, 1)^\top$ in $\mathbb{R}_{\geq 0} \times \mathbb{R}^n$) between $H_*(\mathcal{R}^{\bullet \to 2\bullet}(\mathscr{f}|_P))$ and $\mathrm{Lan}_\iota H_*(\mathscr{f})$.

Since vertical interleavings are stronger than ordinary interleavings (Remark 1), Theorem 6 not only provides approximation guarantees for any interleaving-stable invariant of $\mathrm{Lan}_\iota H_*(\mathscr{f})$ by the corresponding invariant of $H_*(\mathcal{R}^{\bullet \to 2\bullet}(\mathscr{f}|_P))$, but it also offers refined statements for invariants that behave in a particular way under vertical interleavings, such as for instance multigraded Betti numbers as proven in the full version [1] and illustrated in Figure 2.

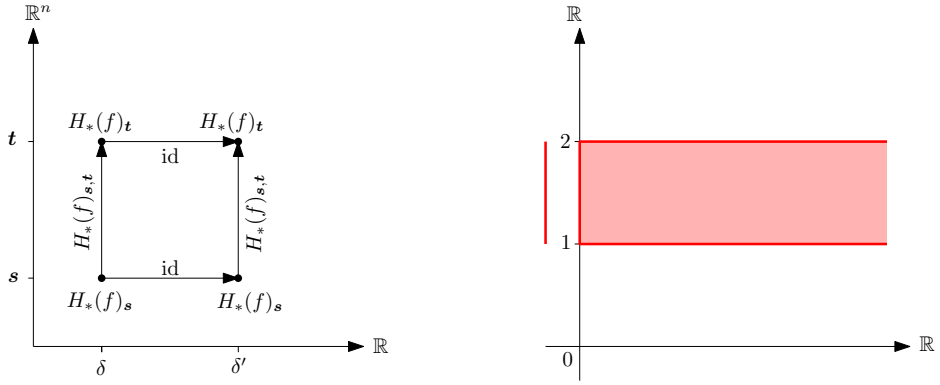


Figure 1 Left: an illustration of the left Kan extension $\mathrm{Lan}_\iota H_*(\mathscr{f}) : \mathbb{R}_{\geq 0} \times \mathbb{R}^n \to \mathbf{vec}$, with identity maps horizontally and the structural morphisms of $H_*(\mathscr{f})$ vertically. **Right**: (case $n = 1$) the left Kan extension of the interval module $\Bbbk^{[1,2]}$ is the interval module $\Bbbk^{\mathbb{R}_{\geq 0} \times [1,2]}$ (in red).

In practice, function values may be corrupted with measurement noise, meanwhile geodesic distances may have to be approximated from the input data. Our framework can account for these imprecisions by suitably adapting the parameters of the estimators and their approximation bounds (see the full version [1, Section 3.3]).

Still for practical purposes, we discuss the computation of our estimators, in particular we provide an algorithm for computing a free presentation of $H_*(\mathcal{R}^{\bullet \to 2\bullet}(\mathscr{f}|_P))$, and thus also of $H_*(\mathcal{R}^{\delta \to 2\delta}(\mathscr{f}|_P))$ for any fixed $\delta \geq 0$, in time that is comparable to that of computing a free presentation of the persistent homology of a single multifiltration. From a free presentation, a variety of invariants, including multigraded Betti numbers, can be derived using existing software. We have implemented our algorithm and used it in our experiments.

Structure of the paper

We provide background material in Section 2. Our problem statement, estimators, and main results are detailed in Section 3. In Section 4 we discuss algorithmic aspects. Finally, in Section 5 we present experimental results that illustrate our theoretical guarantees and we investigate additional properties of our estimators. Further algorithmic details, as well as the proofs of our main results, can be found in the full version [1].

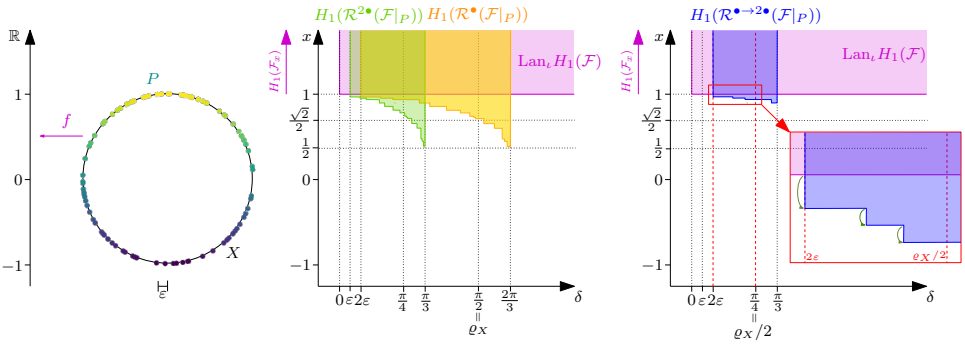


Figure 2 Left: the vertical height function on a sampled unit circle X in the plane. Distances on the circle are given by arclength. **Center**: $H_1(\mathscr{f})$ has a single interval summand (in magenta), starting at height 1, which extends to the free module $\mathrm{Lan}_\iota H_1(\mathscr{f})$ generated at $(0,1)$ in $\mathbb{R}_{\geq 0} \times \mathbb{R}$. The modules $H_1(\mathcal{R}^{\bullet}(\mathscr{f}|_P))$ (in yellow) and $H_1(\mathcal{R}^{2\bullet}(\mathscr{f}|_P))$ (in green) are interval modules in this simple scenario. **Right**: the estimator $H_1(\mathcal{R}^{\bullet\to 2\bullet}(\mathscr{f}|_P))$ (in blue), which is also an interval module, approximates the target $\mathrm{Lan}_\iota H_1(\mathscr{f})$ in the vertical interleaving distance within any slab $[2\varepsilon, \delta_0] \times \mathbb{R}$ with $\delta_0 < \varrho_X/2$, as per Theorem 6. In turn, the vertical interleaving between the two modules implies a vertical matching between their multigraded Betti numbers within the slab (illustrated by green arrows in the close-up view), as per Corollary 3.8 in the full version [1].

2 Background

We assume some familiarity with basic category theory, algebraic topology, measure theory, and topological data analysis. Let **Top** denote the category of topological spaces, and **vec** (resp. **Vec**) the category of finite-dimensional (resp. all) vector spaces over a fixed field $\Bbbk$. We also fix an arbitrary degree in homology, denoted by $*$, and we write $H_*(-)$ for singular homology groups in degree $*$ with coefficients in $\Bbbk$.

Filtrations and persistence modules

We see $\mathbb{R}^n$ as the product of n copies of the real line, equipped with the product order noted $\leq$. Thus, two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ satisfy $\boldsymbol{x} \leq \boldsymbol{y}$ whenever $\boldsymbol{x}_i \leq \boldsymbol{y}_i$ for all $1 \leq i \leq n$. We denote by $(-\infty, \boldsymbol{x}]$ the downset $\{\boldsymbol{y} \in \mathbb{R}^n \mid \boldsymbol{y} \leq \boldsymbol{x}\}$, and by $[\boldsymbol{x}, +\infty)$ the upset $\{\boldsymbol{y} \in \mathbb{R}^n \mid \boldsymbol{y} \geq \boldsymbol{x}\}$. An *$n$-parameter filtration* is a functor $\mathcal{F} : \mathbb{R}^n \to \mathbf{Top}$ whose constituent maps are inclusions, which means that we have a topological space $\mathcal{F}_{\boldsymbol{x}}$ for all $\boldsymbol{x} \in \mathbb{R}^n$ and an inclusion $\mathcal{F}_{\boldsymbol{x},\boldsymbol{y}} \colon \mathcal{F}_{\boldsymbol{x}} \hookrightarrow \mathcal{F}_{\boldsymbol{y}}$ for all $\boldsymbol{x} \leq \boldsymbol{y} \in \mathbb{R}^n$. An *$n$-parameter persistence module* is a functor $M : \mathbb{R}^n \to \mathbf{Vec}$, which means that we have a $\Bbbk$-vector space $M_{\boldsymbol{x}}$ for all $\boldsymbol{x} \in \mathbb{R}^n$ and a $\Bbbk$-linear map $M_{\boldsymbol{x},\boldsymbol{y}} \colon M_{\boldsymbol{x}} \to M_{\boldsymbol{y}}$ for all $\boldsymbol{x} \leq \boldsymbol{y} \in \mathbb{R}^n$, such that $M_{\boldsymbol{x},\boldsymbol{x}} = \mathrm{Id}_{M_{\boldsymbol{x}}}$ and $M_{\boldsymbol{x},\boldsymbol{z}} = M_{\boldsymbol{y},\boldsymbol{z}} \circ M_{\boldsymbol{x},\boldsymbol{y}}$ for all $\boldsymbol{x} \leq \boldsymbol{y} \leq \boldsymbol{z} \in \mathbb{R}^n$. A *morphism* of persistence modules $M \to N$ is a natural transformation between functors, i.e., a family of linear maps $\varphi_{\boldsymbol{x}} \colon M_{\boldsymbol{x}} \to N_{\boldsymbol{x}}$ such that $N_{\boldsymbol{x},\boldsymbol{y}} \circ \varphi_{\boldsymbol{x}} = \varphi_{\boldsymbol{y}} \circ M_{\boldsymbol{x},\boldsymbol{y}}$ for all $\boldsymbol{x} \leq \boldsymbol{y} \in \mathbb{R}^n$.

Geodesic metric spaces

Throughout the paper, unless otherwise specified, (X, d_X) is a compact geodesic metric space. This means that, for all $x, y \in X$, there is a shortest continuous path from x to y in X, of length equal to $d_X(x, y) < \infty$. In particular, the space is path-connected. Let P be a finite set of points of X. Given $\varepsilon > 0$, we say P is a *geodesic ε-sample* of X if $d_{\mathrm{H}}(P, X) < \varepsilon$, where d_{H}

denotes the Hausdorff distance in (X, d_X). Define $B_X(x, r)$ as the open geodesic ball centered at $x \in X$ with radius r, given by $B_X(x, r) = \{x' \in X \mid d_X(x, x') < r\}$. A ball $B_X(x, r)$ is said to be *convex* if, for any pair of points y and y' in $B_X(x, r)$, the shortest path in X that connects y to y' is unique and included in $B_X(x, r)$. Let $\varrho(x) := \inf\{r > 0 \mid B(x, r) \text{ is not convex}\}$, and let $\varrho_X := \inf\{\varrho(x) \mid x \in X\} \geq 0$. This quantity, called the *convexity radius* of X, plays an important role because intersections of convex sets are convex and therefore contractible, so the Nerve Lemma [13, Corollary 4G.3] applies to unions of balls of radii less than ϱ_X.

Functional and geometric filtrations

Given a geodesic metric space (X, d_X) and a function $\ell\colon X \to \mathbb{R}^n$, the *sublevel filtration* $\mathcal{F} : \mathbb{R}^n \to \mathbf{Top}$ of ℓ is defined by $\mathcal{F}_{\boldsymbol{x}} := \ell^{-1}((-\infty, \boldsymbol{x}])$ for all $\boldsymbol{x} \in \mathbb{R}^n$. The *persistent homology* of $\mathcal{F}$ (and, by extension, of ℓ) is the persistence module $H_*(\mathcal{F})$ (also written $H_*(\ell)$) defined by $H_*(\mathcal{F})_{\boldsymbol{x}} := H_*(\mathcal{F}_{\boldsymbol{x}})$ and $H_*(\mathcal{F})_{\boldsymbol{x},\boldsymbol{y}} := H_*(\mathcal{F}_{\boldsymbol{x},\boldsymbol{y}})$ for all $\boldsymbol{x} \leq \boldsymbol{y} \in \mathbb{R}^n$.

Given a finite set P of points in X, and a real parameter $\delta \geq 0$, we call *δ-offset* of P the union of open geodesic balls $\mathcal{O}^\delta(P) := \bigcup_{p\in P} B_X(p, \delta)$. We call *$\delta$-Čech Complex* of P the nerve of the collection of balls $\{B_X(p, \delta) \mid p \in P\}$, defined as the abstract simplicial complex $\mathcal{C}^\delta(P) := \{\emptyset \neq \sigma \subseteq P \mid \bigcap_{p\in\sigma} B_X(p, \delta) \neq \emptyset\}$. For $\delta < \varrho_X$, the balls $B_X(p, \delta)$ and their intersections are either empty or convex, so the Nerve Lemma [13, Corollary 4G.3] ensures that the homology groups $H_*(\mathcal{O}^\delta(P))$ and $H_*(\mathcal{C}^\delta(P))$ are isomorphic. We call *δ-Rips complex* of P the abstract simplicial complex $\mathcal{R}^\delta(P) := \{\emptyset \neq \sigma \subseteq P \mid \max_{p,q\in\sigma} d_X(p, q) < \delta\}$. Čech and Rips complexes of P are related as follows: $\mathcal{C}^\delta(P) \subseteq \mathcal{R}^{2\delta}(P) \subseteq \mathcal{C}^{2\delta}(P)$ for all $\delta \geq 0$.

Interleavings

Let $\boldsymbol{v} \in \mathbb{R}^n_{\geq 0}$ be a fixed vector, and let $\varepsilon \geq 0$. Given any functors $F, G : \mathbb{R}^n \to \mathcal{C}$ to some category $\mathcal{C}$, we say F and G are *$\varepsilon\boldsymbol{v}$-interleaved* if there are two natural transformations $f : F \to G[\varepsilon\boldsymbol{v}]$ and $g : G \to F[\varepsilon\boldsymbol{v}]$ such that $g[\varepsilon\boldsymbol{v}] \circ f = \varphi_F^{2\varepsilon\boldsymbol{v}}$ and $f[\varepsilon\boldsymbol{v}] \circ g = \varphi_G^{2\varepsilon\boldsymbol{v}}$. Here, $-[\varepsilon\boldsymbol{v}]$ denotes the $\varepsilon\boldsymbol{v}$-shift functor, defined on objects by $F[\varepsilon\boldsymbol{v}]_{\boldsymbol{x}} := F_{\boldsymbol{x}+\varepsilon\boldsymbol{v}}$ and on morphisms by $F[\varepsilon\boldsymbol{v}]_{\boldsymbol{x},\boldsymbol{y}} := F_{\boldsymbol{x}+\varepsilon\boldsymbol{v},\boldsymbol{y}+\varepsilon\boldsymbol{v}}$ for all $\boldsymbol{x} \leq \boldsymbol{y} \in \mathbb{R}^n$. Meanwhile, $\varphi_F^{2\varepsilon\boldsymbol{v}}\colon F \to F[2\varepsilon\boldsymbol{v}]$ and $\varphi_G^{2\varepsilon\boldsymbol{v}}\colon G \to G[2\varepsilon\boldsymbol{v}]$ denote the natural morphisms from F and G to their respective $2\varepsilon\boldsymbol{v}$-shifts, defined by $(\varphi_F^{2\varepsilon\boldsymbol{v}})_{\boldsymbol{x}} := F_{\boldsymbol{x},\boldsymbol{x}+2\varepsilon\boldsymbol{v}}$ and $(\varphi_G^{2\varepsilon\boldsymbol{v}})_{\boldsymbol{x}} := G_{\boldsymbol{x},\boldsymbol{x}+2\varepsilon\boldsymbol{v}}$ for all $\boldsymbol{x} \in \mathbb{R}^n$. The *$\boldsymbol{v}$-interleaving distance* is defined as follows:

$$d_{\mathrm{i}}^{\boldsymbol{v}}(F, G) = \inf\{\varepsilon \geq 0 \mid \text{ there exists an } \varepsilon\boldsymbol{v}\text{-interleaving between } F \text{ and } G\} \in [0, +\infty].$$

In the following, we consider two specific types of interleavings:

1. **Ordinary interleavings**: Let $\boldsymbol{v} = \mathbf{1} := (1, \cdots, 1)^T$. Then, $\varepsilon\mathbf{1}$-interleavings are ordinary ε-interleavings from the TDA literature, and $d_{\mathrm{i}}^{\mathbf{1}}$ is called the *ordinary interleaving distance*.
2. **Vertical interleavings**: Let $\boldsymbol{v} = \mathbf{1}_0 := (0, 1, \cdots, 1)^T$. Then, $\varepsilon\mathbf{1}_0$-interleavings are called *vertical ε-interleavings*, and $d_{\mathrm{i}}^{\mathbf{1}_0}$ is called the *vertical interleaving distance*.

▶ Remark 1. Vertical interleaving implies ordinary interleaving. Indeed, the vertical interleaving maps can be composed with the horizontal structure morphisms of the modules to form an ordinary interleaving. The amplitude ε of the interleaving does not change in the process. Therefore, vertical interleaving is a stronger condition than ordinary interleaving.

3 Problem statement and main results

Let (X, d_X) be a compact geodesic metric space with positive convexity radius ϱ_X, and let $\ell : X \to \mathbb{R}^n$ be a continuous function. Here, $\mathbb{R}^n$ is equipped with the ∞-norm. Both X and ℓ are unknown. By the Heine–Cantor Theorem, ℓ admits a modulus of continuity

$\omega\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. Neither ω nor ϱ_X need to be known, as they only play a role in our bounds and not in our constructions. The function $\boldsymbol{f}$ is not assumed to be pointwise finite-dimensional (i.e., $\dim H_*(\boldsymbol{f})_{\boldsymbol{x}}$ finite for every $\boldsymbol{x} \in \mathbb{R}^n$), however our assumptions imply a form of tameness for the persistence module $H_*(\boldsymbol{f})$ (see Corollary 4).

Our input is a finite point cloud $P \subseteq X$ such that $\boldsymbol{f}|_P$, the restriction of $\boldsymbol{f}$ to P, is known, as well as the pairwise geodesic distances $d_X(p,q)$ between the points $p, q \in P$. The problem we address is that of estimating $H_*(\boldsymbol{f})$, the persistent homology of the sublevel filtration of $\boldsymbol{f}$, from our input. For this we assume that P is a geodesic ε-sample of X, for some possibly unknown ε that we assume to be small enough compared to the convexity radius ϱ_X.

To build our estimator, we first construct the *function-Rips multifiltration* $\mathcal{R}^\bullet(\boldsymbol{f}|_P) := \left\{\mathcal{R}^\delta\left(\mathcal{F}_{\boldsymbol{x}} \cap P\right)\right\}_{\delta\in\mathbb{R}_{\geq 0},\, \boldsymbol{x}\in\mathbb{R}^n}$ from the restriction of $\boldsymbol{f}$ to P. Then, following [10], our estimator is a smoothed version of $H_*(\mathcal{R}^\bullet(\boldsymbol{f}|_P))$, denoted by $H_*(\mathcal{R}^{\bullet\to 2\bullet}(\boldsymbol{f}|_P))$ and defined formally as the image of the morphism of persistence modules induced in homology by the inclusion of the filtration $\mathcal{R}^\bullet(\boldsymbol{f}|_P)$ into its horizontal rescaling by a factor of 2, that is:

$$H_*(\mathcal{R}^{\bullet\to 2\bullet}(\boldsymbol{f}|_P)) := \operatorname{Im}\ H_*\left(\mathcal{R}^\bullet(\boldsymbol{f}|_P) \hookrightarrow \mathcal{R}^{2\bullet}(\boldsymbol{f}|_P)\right),$$

where $\mathcal{R}^{2\bullet}(\boldsymbol{f}|_P) := \left\{\mathcal{R}^{2\delta}(\mathcal{F}_{\boldsymbol{x}} \cap P)\right\}_{\delta\in\mathbb{R}_{\geq 0},\, \boldsymbol{x}\in\mathbb{R}^n}$.

▶ Remark 2. It is known that a single Rips complex of a finite sampling can recover the homological type of a sufficiently regular underlying continuous space [14, 16]. Yet, in full generality it takes a pair of Rips complexes with parameters within a factor of at least 2 of each other to do so [7]. Similarly, it was recently shown that $H_*(\mathcal{R}^\bullet(\boldsymbol{f}|_P))$ itself can approximate $H_*(\boldsymbol{f})$ provided X has curvature bounded above locally [22]. Yet, it is unknown whether the same holds in our more general setting, which is why we use $H_*(\mathcal{R}^{\bullet\to 2\bullet}(\boldsymbol{f}|_P))$.

In the course of our analysis of the behavior of $H_*(\mathcal{R}^{\bullet\to 2\bullet}(\boldsymbol{f}|_P))$, we will consider intermediate constructions, namely the persistent homologies of:

- the *function-offset filtration* $\mathcal{O}^\bullet(\boldsymbol{f}|_P) := \left\{\mathcal{O}^\delta(\mathcal{F}_{\boldsymbol{x}} \cap P)\right\}_{\delta\in\mathbb{R}_{\geq 0},\, \boldsymbol{x}\in\mathbb{R}^n}$;
- the *function-Čech filtration* $\mathcal{C}^\bullet(\boldsymbol{f}|_P) := \left\{\mathcal{C}^\delta(\mathcal{F}_{\boldsymbol{x}} \cap P)\right\}_{\delta\in\mathbb{R}_{\geq 0},\, \boldsymbol{x}\in\mathbb{R}^n}$.

These can also serve as alternative estimators when they can be built, for instance – in the case of $\mathcal{C}^\bullet(\boldsymbol{f}|_P)$ – when we can test whether finitely many open geodesic balls intersect.

We use our estimators in two types of scenarios: (1) when we have an estimate of the value ε for which P is an ε-sample of X, and (2) when we do not have such an estimate.

3.1 Estimating $H_*(\boldsymbol{f})$ with known ε

Suppose we know ε or some reasonable estimate. Then we can approximate our target $H_*(\boldsymbol{f})$ by restricting our estimators to a vertical hyperplane $\{\delta\} \times \mathbb{R}^n \subseteq \mathbb{R}_{\geq 0} \times \mathbb{R}^n$ for a suitable choice of parameter δ. We denote these restrictions respectively by $H_*(\mathcal{O}^\delta(\boldsymbol{f}|_P))$, $H_*(\mathcal{C}^\delta(\boldsymbol{f}|_P))$ and $H_*(\mathcal{R}^{\delta\to 2\delta}(\boldsymbol{f}|_P))$.

▶ **Theorem 3.** *Let X be a compact geodesic space, let $\boldsymbol{f} : X \to \mathbb{R}^n$ be an ω-continuous function for some modulus of continuity $\omega : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, and let P be a finite geodesic ε-sample of X. Then:*

(i) *for $\forall \delta \geq \varepsilon$, the modules $H_*(\boldsymbol{f})$ and $H_*(\mathcal{O}^\delta(\boldsymbol{f}|_P))$ are ordinarily $\omega(\delta)$-interleaved;*

(ii) *for $\forall \delta \in [\varepsilon, \varrho_X)$, the modules $H_*(\boldsymbol{f})$ and $H_*(\mathcal{C}^\delta(\boldsymbol{f}|_P))$ are ordinarily $\omega(\delta)$-interleaved;*

(iii) *for $\forall \delta \in [2\varepsilon, \frac{\varrho_X}{2})$, modules $H_*(\boldsymbol{f})$ and $H_*(\mathcal{R}^{\delta\to 2\delta}(\boldsymbol{f}|_P))$ are ordinarily $\omega(2\delta)$-interleaved.*

The proof of this result is given in the full version [1, Section 4]. The conclusion still holds under noisy input function values or geodesic distances, as detailed in [1, Section 3.3].

Theorem 3 can be read in two different ways. First, when the modulus of continuity ω is known, we obtain explicit bounds on the interleaving distance between our estimators and the target $H_*(f)$. Second, when we only know that f is continuous, the Heine–Cantor Theorem implies that f admits some modulus of continuity ω, and even without knowning that ω, we are guaranteed by Theorem 3 that our estimators converge to $H_*(f)$ as ε goes to zero under suitable choices of parameter δ.

When ε is known exactly, the best choice for δ is the lower bound of its admissible interval, i.e., $\delta = \varepsilon$ in Items i and ii and $\delta = 2\varepsilon$ in Item iii. When ε is only known approximately, the admissible interval provides some leeway for the choice of δ.

An immediate consequence of Theorem 3 is that $d_{\mathrm{i}}^{\mathbf{1}}$-stable invariants computed on our estimators approximate the corresponding invariants defined on the target $H_*(f)$. For instance, the approximation guarantee obtained for the multigraded Betti numbers of $H_*(\mathcal{R}^{\delta\to 2\delta}(f|_P))$ is given in the full version [1, Corollary 3.3].

Another consequence of Theorem 3 (i) is that, under our assumptions, the persistence module $H_*(f)$ satisfies the following form of tameness, as proven in the full version [1].

▶ **Corollary 4.** *Let X, ω, f be as in Theorem 3. Then the module $H_*(f)$ is tame in the sense that the rank $\mathrm{rk}(H_*(f)_{\boldsymbol{x},\boldsymbol{y}})$ is finite for all $\boldsymbol{x} < \boldsymbol{y} \in \mathbb{R}^n$ with $\boldsymbol{x}_i < \boldsymbol{y}_i$ for each $i \in \{1, \cdots, n\}$.*

Suppose now the points of P are i.i.d. samples drawn from some unknown probability measure μ supported on X. In this setting, standard statistical techniques can be used to estimate ε with high probability, then Theorem 3 can be applied to get statistical estimates of $H_*(f)$ or of any $d_{\mathrm{i}}^{\mathbf{1}}$-stable invariant thereof. We summarize our results in Theorem 5. The relevant statistical background and definitions are provided in the full version [1, Section 2], together with the analysis of the statistical estimators [1, Section 6].

▶ **Theorem 5.** *Let X, ω and f be as in Theorem 3. Assume that the convexity radius satisfies $\varrho_X > 0$, that the modulus of continuity ω has a controlled vanishing rate $\delta = O(\omega(\delta))$ as $\delta \to 0$, and that the points $p_1, \ldots, p_k$ of P are i.i.d. samples drawn from some unknown (a,b)-standard probability measure μ whose support is the entire space X.*

- **(i)** *Suppose a, b are known. Let $\delta_k := 4\left(\frac{2\log k}{ak}\right)^{1/b}$. Then, the sequences of approximations $\left(H_*(\mathcal{O}^{\delta_k}(f|_P))\right)_{k\in\mathbb{N}}$, $\left(H_*(\mathcal{C}^{\delta_k}(f|_P))\right)_{k\in\mathbb{N}}$ and $\left(H_*(\mathcal{R}^{\delta_k\to 2\delta_k}(f|_P))\right)_{k\in\mathbb{N}}$ of $H_*(f)$ are ω-quasi-minimax consistent estimators of $H_*(f)$ in the following sense: their expected interleaving distances to $H_*(f)$ are at most a constant times $\omega\left(8\left(\frac{2\log k}{ak}\right)^{1/b}\right)$; meanwhile, in the worst case over n, X, f as above, for any other estimator of $H_*(f)$, the supremum over all (a,b)-standard measures μ of the expected interleaving distance between the estimator and its target $H_*(f)$ is at least a constant times $\omega\left(\frac{1}{2}\left(\frac{1}{ak}\right)^{1/b}\right)$.*
- **(ii)** *Suppose a, b are unknown, and assume further that X is a compact smooth manifold and that the measure μ decomposes as $\mu = \mu_1 + \mu_2$, where $\mu_2(X) > 0$ and μ_2 is absolutely continuous with respect to the uniform measure on X, with a positive density on X. Then, there is an explicit sequence of random positive numbers $(\hat{\delta}_k)_{k\in\mathbb{N}}$ such that the induced plug-in estimators $\left(H_*(\mathcal{O}^{\hat{\delta}_k}(f|_P))\right)_{k\in\mathbb{N}}$, $\left(H_*(\mathcal{C}^{\hat{\delta}_k}(f|_P))\right)_{k\in\mathbb{N}}$, and $\left(H_*(\mathcal{R}^{\hat{\delta}_k\to 2\hat{\delta}_k}(f|_P))\right)_{k\in\mathbb{N}}$ are ω-quasi-minimax consistent estimators of $H_*(f)$.*

3.2 Estimating $H_*(f)$ with unknown ε

When ε is unknown and cannot be effectively estimated – for instance when the assumptions in Theorem 5 are not satisfied, we adopt the traditional approach in persistence theory which is to not fix the scale parameter δ but rather to consider it as an extra filtration

parameter. This means using the full $(n+1)$-parameter modules $H_*(\mathcal{O}^\bullet(f|_P))$, $H_*(\mathcal{C}^\bullet(f|_P))$ and $H_*(\mathcal{R}^{\bullet\to 2\bullet}(f|_P))$ as estimators. In order to state approximation results, we extend our target $H_*(f)$ to a persistence module over $\mathbb{R}_{\geq 0}\times\mathbb{R}^n$ by taking its left Kan extension $\mathrm{Lan}_\iota H_*(f)$ along the poset embedding $\iota:\mathbb{R}^n\hookrightarrow\mathbb{R}_{\geq 0}\times\mathbb{R}^n$ given by $\boldsymbol{x}\mapsto(0,\boldsymbol{x})$ for all $\boldsymbol{x}\in\mathbb{R}^n$. Since the category $\mathbb{R}^n$ is small and the category **Vec** is cocomplete, $\mathrm{Lan}_\iota H_*(f)$ is well-defined as a functor $\mathbb{R}_{\geq 0}\times\mathbb{R}^n\to\mathbf{Vec}$, and given pointwise by the colimit formula [19]:

$$\forall(\delta,\boldsymbol{x})\in\mathbb{R}_{\geq 0}\times\mathbb{R}^n,\quad \mathrm{Lan}_\iota H_*(f)_{(\delta,\boldsymbol{x})}=\varinjlim_{\substack{\boldsymbol{y}\in\mathbb{R}^n\\ \iota(\boldsymbol{y})\leq(\delta,\boldsymbol{x})}} H_*(f)_{\boldsymbol{y}}\cong H_*(f)_{\boldsymbol{x}}. \tag{1}$$

Moreover, $\mathrm{Lan}_\iota H_*(f)$ has structural morphisms that are identity maps horizontally (i.e., along the first coordinate axis) and the structural morphisms of $H_*(f)$ vertically (i.e., along any direction orthogonal to the first coordinate axis), as illustrated in Figure 1.

▶ **Theorem 6.** *Let X, ω, f, P and ε as in Theorem 3. Then:*

- **(i)** *for $\forall\delta_0\geq\varepsilon$, within the slab $[\varepsilon,\delta_0]\times\mathbb{R}^n$, the restricted modules $\mathrm{Lan}_\iota H_*(f)|_{[\varepsilon,\delta_0]\times\mathbb{R}^n}$ and $H_*(\mathcal{O}^\bullet(f|_P))|_{[\varepsilon,\delta_0]\times\mathbb{R}^n}$ are vertically $\omega(\delta_0)$-interleaved;*
- **(ii)** *for $\forall\delta_0\in[\varepsilon,\varrho_X)$, within the slab $[\varepsilon,\delta_0]\times\mathbb{R}^n$, the restricted modules $\mathrm{Lan}_\iota H_*(f)|_{[\varepsilon,\delta_0]\times\mathbb{R}^n}$ and $H_*(\mathcal{C}^\bullet(f|_P))|_{[\varepsilon,\delta_0]\times\mathbb{R}^n}$ are vertically $\omega(\delta_0)$-interleaved;*
- **(iii)** *for $\forall\delta_0\in[2\varepsilon,\varrho_X/2)$, within the slab $[2\varepsilon,\delta_0]\times\mathbb{R}^n$, the restricted modules $\mathrm{Lan}_\iota H_*(f)|_{[2\varepsilon,\delta_0]\times\mathbb{R}^n}$ and $H_*(\mathcal{R}^{\bullet\to 2\bullet}(f|_P))|_{[2\varepsilon,\delta_0]\times\mathbb{R}^n}$ are vertically $\omega(2\delta_0)$-interleaved.*

The proof is given in the full version [1, Section 5]. Similar to Theorem 3, the conclusion still holds under noisy input function values or geodesic distances – see [1, Section 3.3].

This approximation result encapsulates the previous one in the sense that restricting the estimators to the vertical hyperplane $\{\delta\}\times\mathbb{R}^n$ for $\delta=\delta_0$ recovers Theorem 3. However, the result says something deeper, namely: that the interleavings in the vertical hyperplanes $\{\delta\}\times\mathbb{R}^n$, for δ ranging over $[\varepsilon,\delta_0]$ (resp. $[2\varepsilon,\delta_0]$), commute with the horizontal structural morphisms of the modules, so that they all together form a vertical interleaving in the slab $[\varepsilon,\delta_0]\times\mathbb{R}^n$ (resp. $[2\varepsilon,\delta_0]\times\mathbb{R}^n$). Intuitively, this means that prominent features in $H_*(f)$ are not just present in $H_*(\mathcal{R}^{\delta\to 2\delta}(f|_P))$ at individual scales δ in the range $[2\varepsilon,\delta_0]$, but that they persist across large ranges of scales in $H_*(\mathcal{R}^{\bullet\to 2\bullet}(f|_P))$. In particular, they are not ephemeral in $\mathbb{R}_{\geq 0}\times\mathbb{R}^n$ and can be captured by stable invariants.

In terms of approximation accuracy, the result says that, the more one looks to the left (i.e., toward small positive values of δ_0) in the parameter space $\mathbb{R}_{\geq 0}\times\mathbb{R}^n$, the more precisely the estimators approximate the target, until some point where δ_0 becomes too small and the approximation breaks down. On the opposite side, the more one looks to the right (i.e., toward large values of δ_0), the more the estimators drift away from the target, until again the approximation eventually breaks down if the estimator is $H_*(\mathcal{C}^\bullet(f|_P))$ or $H_*(\mathcal{R}^{\bullet\to 2\bullet}(f|_P))$. See Figure 2 for an illustration, and Figure 3 for a comparison of Theorem 6 with Theorem 3.

Theorem 6, like its counterpart Theorem 3, implies that provably $d_{\mathrm{i}}^{\mathbf{1}}$-stable invariants computed from our estimators approximate the corresponding invariants defined on their target, which this time is $\mathrm{Lan}_\iota H_*(f)$. And since the interleaving is now vertical – hence stronger than an ordinary interleaving according to Remark 1, we can get more precise statements for some invariants. For instance, the induced bottleneck matching on multigraded Betti numbers is vertical, as proven in the full version [1] and illustrated in Figure 2.

4 Estimators Computation

Our main estimators, $H_*(\mathcal{R}^{\delta\to 2\delta}(f|_P))$ and $H_*(\mathcal{R}^{\bullet\to 2\bullet}(f|_P))$, derive from the function Rips filtration $\mathcal{R}^\bullet(f|_P)$, whose computation is standard – see [1, Section 7] for details. Here we focus on the computation of invariants once $\mathcal{R}^\bullet(f|_P)$ has been built. The difficulty

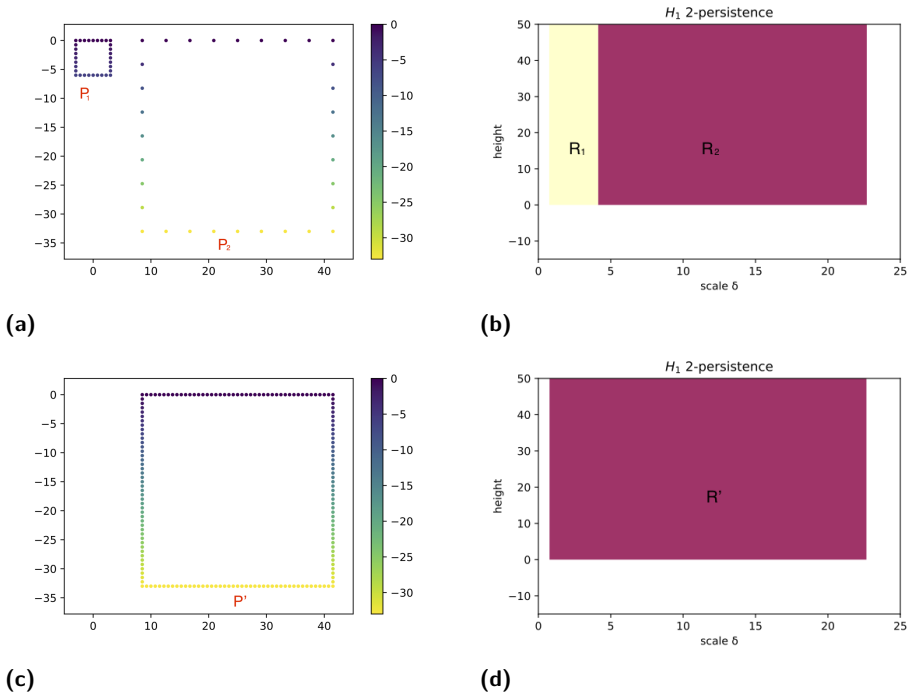


Figure 3 Contrasting Theorem 6 with Theorem 3. **(a)**: The input is $P = P_1 \sqcup P_2$, where P_1 and P_2 are two point clouds regularly sampled from two disjoint squares X_1, X_2 in the plane. Distances within each square are shortest-path distances along the boundary, while distances between squares are infinite. Here ℓ is the vertical height function, and its persistent homology in degree 1 is considered. **(b)**: The estimator $H_1(\mathcal{R}^{\bullet\to 2\bullet}(\ell|_P))$ is an interval-decomposable module whose summands have half-open rectangle supports, respectively R_1 (in yellow) and R_2 (in red). The scalings of X_1, X_2, and their respective sampling densities, have been adjusted so that $R_1 \cap R_2 = \emptyset$ while $R_1 \cup R_2$ is a single rectangle R' shown in subfigure **(d)**. In turn, the interval module with support R' can be realized as the degree 1 persistent homology of another sample P' from a square in the plane, shown in subfigure **(c)**. Theorem 6 guarantees that $H_1(\mathcal{R}^{\bullet\to 2\bullet}(\ell|_P))$ is interleaved with $\mathrm{Lan}_\iota H_1(\ell|_{X_1})$ over R_1 and with $\mathrm{Lan}_\iota H_1(\ell|_{X_2})$ over R_2, leading to the two summands in **(b)**. By contrast, Theorem 3 only guarantees interleavings within the vertical slices, which is not sufficient to discriminate the module with two summands in **(b)** from the module with a single summand in **(d)**.

is that our estimators do not come from a single filtration, but from the inclusion of two nested filtrations sitting inside $\mathcal{R}^\bullet(\ell|_P)$. For some invariants, such as the Hilbert function or rank invariant, existing algorithms adapt straightforwardly [5]. However, for minimal free presentations – from which many invariants, including multigraded Betti numbers, derive – there is no efficient existing algorithm, except in special situations such as when homology in degree 0 is considered, see the full version [1, Section 7.2.3] for details.

To address this, we provide an algorithm for the general problem of computing a free presentation of the image of any morphism of finitely presented persistence modules $\tau : M \to N$, and then we apply it to the special case of $H_*(\mathcal{R}^{\bullet\to 2\bullet}(\ell|_P))$. Here we merely outline the mathematical construction of a free presentation of $\mathrm{Im}(\tau)$, which explains our approach. The algorithmic details and pseudocode are provided in [1, Section 7]. Note that a simpler variant to obtain a free presentation of the cokernel of τ appeared in [12, Proposition 4.14].

Let $\tau : M \to N$ be a morphism between finitely presented persistence modules, with respective free presentations $p_1 : P_1 \to P_0$ and $q_1 : Q_1 \to Q_0$. Since P_0 is projective, we can lift τ to a map $\gamma : P_0 \to Q_0$ so that the following diagram commutes:

$$\begin{array}{ccccccccc} P_1 & \xrightarrow{p_1} & P_0 & \xrightarrow{p_0} & M & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle \gamma} & & \downarrow{\scriptstyle \tau} & & \\ Q_1 & \xrightarrow{q_1} & Q_0 & \xrightarrow{q_0} & N & \longrightarrow & 0. \end{array} \tag{2}$$

From this diagram, we can derive a free presentation of $\mathrm{Im}(\tau)$ as follows. Consider the pullback P_1' of the subdiagram $P_0 \xrightarrow{\gamma} Q_0 \xleftarrow{q_1} Q_1$, i.e., the kernel of the morphism $P_0 \oplus Q_1 \xrightarrow{\gamma - q_1} Q_0$, where $(\gamma - q_1)(u, v) := \gamma(u) - q_1(v)$ for any $(u, v) \in P_0 \oplus Q_1$. Let $\pi_0 : P_1' \to P_0$ and $\pi_1 : P_1' \to Q_1$ denote the canonical projections from $P_1' \subseteq P_0 \oplus Q_1$ to P_0 and Q_1, respectively. Next, let P_2' be the free cover of P_1', yielding the surjection $\beta : P_2' \twoheadrightarrow P_1'$. We illustrate these morphisms in the following commutative diagram, where P_1', P_0, Q_0, Q_1 form a pullback square:

$$\begin{array}{ccccccccc} P_2' & \xrightarrow{\beta} & P_1' & \xrightarrow{\pi_0} & & & & & \\ & \searrow^{\pi_1} & & & & & & & \\ & & & & P_1 & \xrightarrow{p_1} & P_0 & \xrightarrow{p_0} & M & \longrightarrow & 0 \\ & & & & & & \downarrow{\scriptstyle \gamma} & & \downarrow{\scriptstyle \tau} & & \\ & & & & Q_1 & \xrightarrow{q_1} & Q_0 & \xrightarrow{q_0} & N & \longrightarrow & 0 \end{array} \tag{3}$$

This diagram yields a free presentation of $\mathrm{Im}(\tau)$ as follows (proof is in the full version [1]):

▶ **Proposition 7.** *The sequence $P_2' \xrightarrow{\pi_0 \circ \beta} P_0 \xrightarrow{\eta} \mathrm{Im}(\tau) \to 0$ is exact, where η comes from the factorization of $\tau \circ p_0$ through its image $\mathrm{Im}(\tau)$. Consequently, the morphism $\pi_0 \circ \beta : P_2' \to P_0$ is a free presentation of $\mathrm{Im}(\tau)$, and in fact a finite free presentation.*

5 Experiments

In this section, we illustrate our approximation and convergence results through practical experiments on both synthetic and real-world datasets.

5.1 Toy examples

We validate the claims in Theorems 3 and 6 through a simple example consisting of two circles with distinct radii and concentration levels, as shown in Figure 4 (a). In Figure 4 (b) we analyze the error between the estimator $H_1(\mathcal{R}^{\delta \to 2\delta}(f|_P))$ and the target $H_1(f)$ as a function of the scale parameter δ. As guaranteed by Theorem 3, for δ within the range $[2\varepsilon, \varrho_X/2)$ the error between the estimator and the target is small, controlled by δ. The approximation remains good even for δ going beyond that range, which is not explained by the theory but occurs here because the space and function are very regular in this example.

In Figure 4 (c) we show a visualization of our estimator $H_1(\mathcal{R}^{\bullet \to 2\bullet}(f|_P))$ using `MMA` [17]. As expected from Theorem 6, we observe two topological features in our estimator, corresponding to the two circles in the dataset. The values of δ at which these features appear then disappear are aligned with the endpoints of the interval of values of δ for which $H_1(\mathcal{R}^{\delta \to 2\delta}(f|_P))$ is a good approximation of $H_1(f)$ in Figure 4 (b).

For comparison, in Figure 4 (d) we show a visualization of the estimator $H_1(\mathcal{C}^{\bullet}(f|_P))$ using `MMA`, for which we use the Euclidean distance instead of the geodesic distance in order to enable the construction of the function–Čech filtration. Again, we see two topological features corresponding to the two circles, and $H_1(\mathcal{C}^{\delta}(f|_P))$ is a good approximation of $H_1(f)$ for δ within a fairly large range.

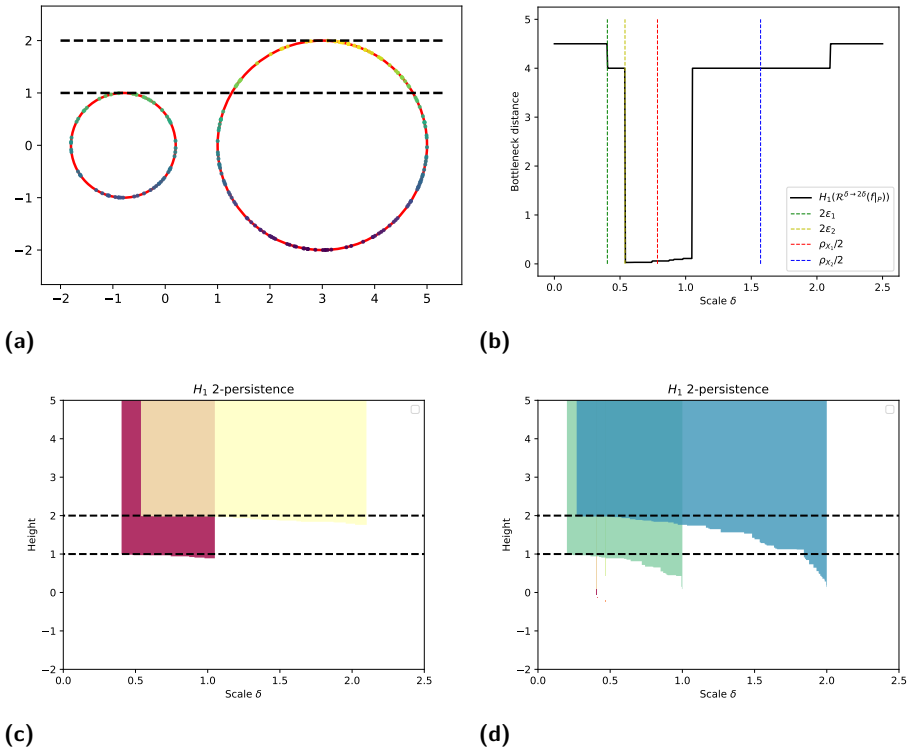


Figure 4 (a): A sample P from a space $X = X_1 \sqcup X_2$ composed of two circles equipped with geodesic distances, each uniformly sampled with distinct radii and concentration levels. The target function $f\colon X \to \mathbb{R}$ is the height function on the two circles. The barcode of the target $H_1(f)$ consists of two infinite bars, each originating from a level marked as a dashed line. **(b)**: Bottleneck distance between the barcode of the estimator $H_1\left(\mathcal{R}^{\delta\to 2\delta}(f|_P)\right)$ and that of the target $H_1\left(f\right)$ as a function of δ. Here, ε_1 (resp. ε_2) is the sampling error of X_1 (resp. X_2), and ϱ_{X_1} (resp. ϱ_{X_2}) the convexity radius of X_1 (resp. X_2). All infinite bars are truncated at 10 to ensure the bottleneck distances are finite. **(c)**: Visualization of the estimator $H_1\left(\mathcal{R}^{\bullet\to 2\bullet}(f|_P)\right)$ computed using `MMA`. **(d)**: Visualization of the estimator $H_1\left(\mathcal{C}^{\bullet}(f|_P)\right)$ computed using `MMA`. Each colored region represents a persistent topological feature. Dashed lines indicate birth times of bars in the barcode of $H_1(f)$.

We now add noise to the dataset sampled from the two circles, as shown in Figure 5 (a). As proven in the full version [1, Section 3.3], the trend is the same. While the theory requires a factor $\alpha > 2$ in $H_1(\mathcal{R}^{\bullet\to\alpha\bullet}(f|_P))$ to handle metric noise, Figure 5 shows that, in this example, the estimator $H_1(\mathcal{R}^{\bullet\to 2\bullet}(f|_P))$ performs well even with $\alpha = 2$.

5.2 Real dataset of immune cells (involving Lipschitz functions)

In this section, we consider the digitized immunohistochemistry dataset from [24]. This dataset is comprised of manually annotated cell locations from three cell types, namely cytotoxic T lymphocytes, regulatory T lymphocytes, and macrophages, each characterized by the expression of one of the proteins CD8, FoxP3, and CD68, respectively – see Figure 6.

The index $i \in \{1, 2, 3\}$ refers to the protein in the ordered set $\{\text{CD8}, \text{FoxP3}, \text{CD68}\}$, and $f_i\colon [0,1]^2 \to \mathbb{R}$ denotes the negative density estimate associated with that protein. Since the functions f_i are unknown, we estimate the density of each point cloud using kernel density

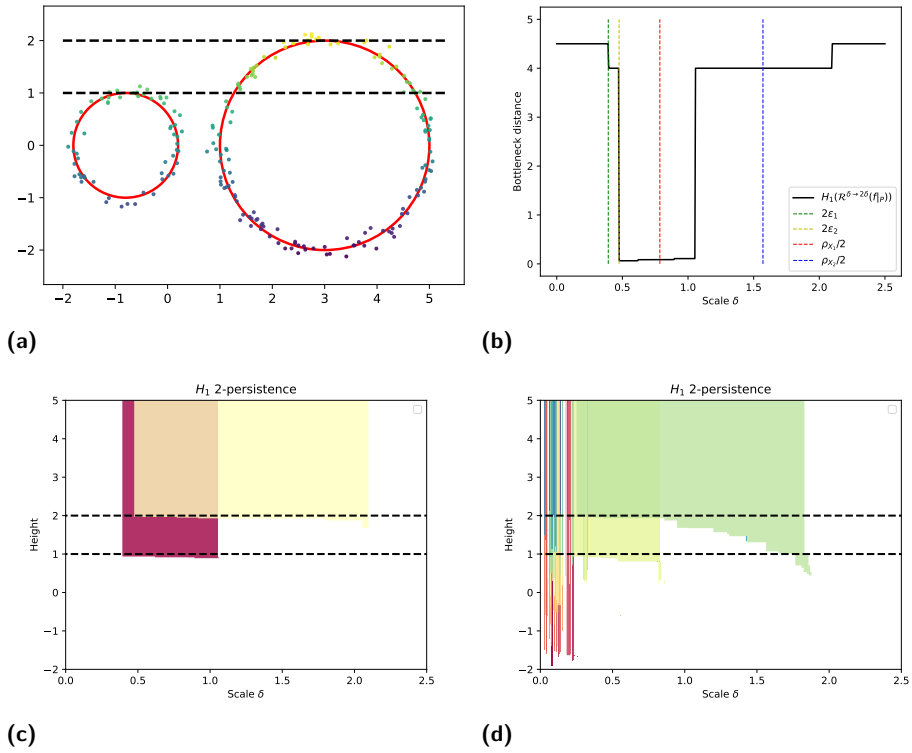


Figure 5 A noisy analog of Figure 4.

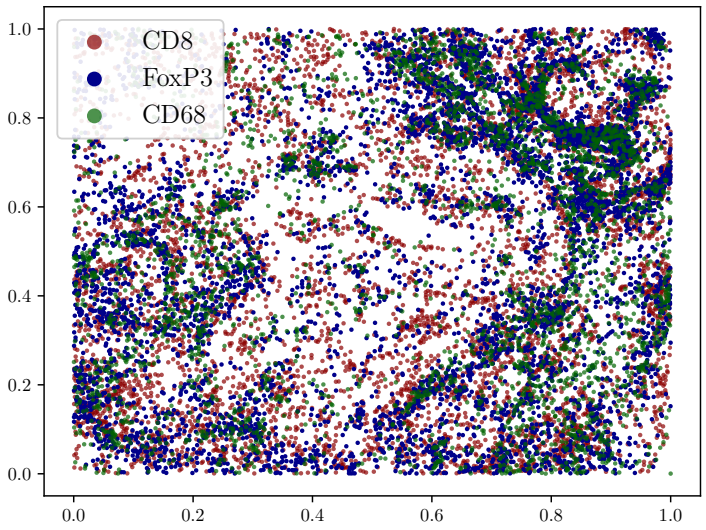
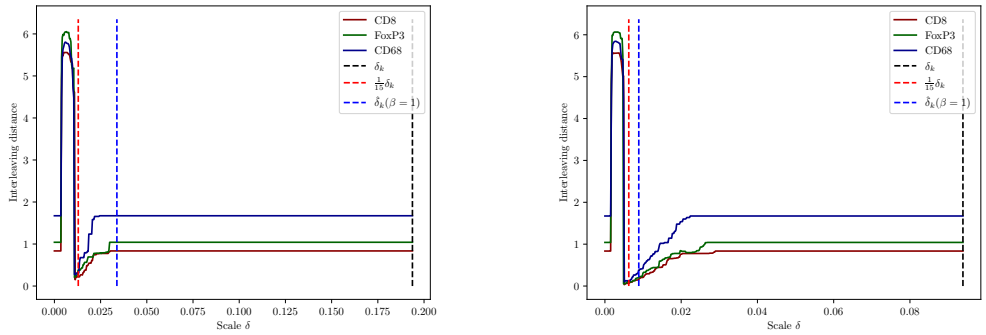


Figure 6 Point cloud dataset, with colors indicating the protein labels CD8, FoxP3, and CD68.



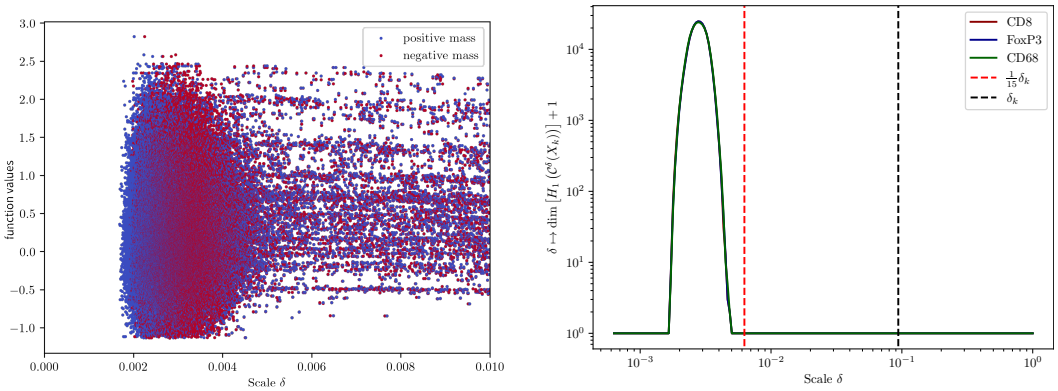
(a) Errors for $H_1(\mathcal{C}^\delta(\mathscr{f}_i|_{X_{10\ 000}}))$.

(b) Errors for $H_1(\mathcal{C}^\delta(\mathscr{f}_i|_{X_{50\ 000}}))$.

Figure 7 Interleaving distance between our estimators and the target $H_1(\mathscr{f})$ w.r.t. the scale δ.

estimation, which yields approximations – also called $\mathscr{f}_1, \mathscr{f}_2, \mathscr{f}_3$ for simplicity – that are Lipschitz continuous by construction. We then discretize the domain $[0,1]^2$ on a grid (using a resolution of $1\,000 \times 1\,000$ for Section 5.2.1 and of 500×500 for Section 5.2.2) and compute the persistence module of the cubical complex induced by the estimated function on this grid, which serves as a proxy for the ground-truth target.

In the following experiments, in order to reach smaller Hausdorff distances between the sample X_k and the domain $X = [0,1]^2$, we first draw 10,000 points uniformly from X, and then select X_k using k-means farthest point sampling applied to this set. In practice, the choice of the rate δ_k prescribed in Theorem 5 can be improved significantly when the sample X_k is well-behaved. We propose heuristics to estimate good rates in non-asymptotic regimes.



(a) Hilbert signed bars of $H_1(\mathcal{C}^\bullet(\mathscr{f}_1|_{X_k}))$.

(b) Pointwise dimension of $H_1(\mathcal{C}^\delta(X_k))$ w.r.t δ.

Figure 8 Heuristic to determine a good multiplicative constant for the rate δ_k, where $k = 50\,000$.

5.2.1 Estimating a single function $\mathscr{f}_i$

We empirically validate our theoretical results by examining the estimators $H_1(\mathcal{C}^\delta(\mathscr{f}_i|_{X_k}))$ with $i \in \{1,2,3\}$ built from the three associated point clouds. We first focus on the convergence rate of the estimators as the sample size k increases, with respect to the choice of scale parameter δ_k. From Theorem 5, since the sampling measure is known – namely the uniform measure on $[0,1]^2$, which is (a,b)-standard with parameters $a = \frac{\pi}{4}$ and $b = 2$ – we

can derive an asymptotic theoretical rate $\delta_k = 4\left(\frac{8\log(k)}{\pi k}\right)^{\frac{1}{2}}$. However, as shown in Figure 7, this choice of δ_k is very conservative, and consequently the estimators $H_1(\mathcal{C}^{\delta_k}(\ell_i|_{X_k}))$ do not perform better than the zero module. This is explained by the fact that δ_k is designed for worst-case sampling measures and asymptotic regimes. As shown in Figure 7, the estimators $H_1(\mathcal{C}^{\delta_k}(\ell_i|_{X_k}))$ achieve good performance only when δ_k is smaller than 0.02, which would require sample sizes k larger than 10^6.

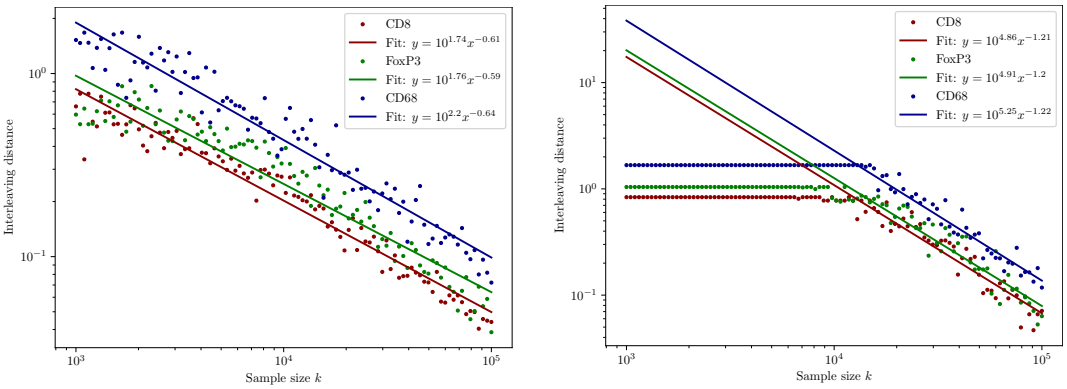
Another option is to consider the estimated rate $\hat{\delta}_k$ mentioned in Theorem 5, defined as $\hat{\delta}_k := d_{\mathrm{H}}(X_{s_k}, X_k)$ where $X_{s_k} \subseteq X_k$ consists of the first s_k sample points of X_k. In Figure 7, we observe that this rate is far less conservative than δ_k and already exhibits its asymptotic behavior when k is larger than 50 000, achieving good performance. In this example though, the rate $\hat{\delta}_k$ has some limitations:

- the bad behavior of the corresponding estimators on small sample sizes ($k \leq 10^4$);
- the dependence of $\hat{\delta}_k$ on a parameter β, which needs to be tuned;
- the fact that the knowledge of the ambient dimension and the probability measure from which the points are sampled are not leveraged.

Here is a heuristic we apply to select a relevant rate δ'_k. Since the dimension of the domain is known, we set δ'_k to be a fixed fraction of the theoretical rate δ_k. This choice is motivated in the full version [1] and ensures that, up to a multiplicative constant, the rate $(1/k)^{1/d}$ cannot be improved for any (possibly deterministic) k-sample of a d-dimensional space. To determine a good multiplicative constant, we notice from Figure 7 that, when $\delta \in [0, d_{\mathrm{H}}(X_k, X))$, the error of the estimators $H_1(\mathcal{C}^{\delta}(\ell_i|_{X_k}))$ soars, due to the fact that $\mathcal{C}^{\delta}(X_k)$ and X do not have the same topology. According to Theorem 3, the optimal scale δ is then given as the smallest value lying above that interval, that is: $d_{\mathrm{H}}(X_k, X)$. Our rate δ'_k is an estimate of this value. In order to compute it, we look at the topological changes of $\mathcal{C}^{\delta}(X_k)$ as δ increases. Since the measure is uniform and thus does not exhibit pathological local behaviors, the artifacts in the topological type of $\mathcal{C}^{\delta}(X_k)$ should all vanish roughly at the same scale, and slightly before $d_{\mathrm{H}}(X_k, X)$. This phenomenon, suggested by the shape of the curves in Figure 7, is further illustrated in Figure 8(a), where we consider the *Hilbert function signed barcode* [18, 21] of $H_1(\mathcal{C}^{\bullet}(\ell_1|_{X_k}))$: after an initial transition phase ($\delta < 0.005$) during which a concentration of positive generators appears, followed then by a concentration of negative generators, the structure of the module stabilizes as these positive and negative generators cancel each other out. This is highlighted when summing the signed generators within the half-plane $(-\infty, \delta] \times \mathbb{R}$, or equivalently, when computing the pointwise dimension of $H_1(\mathcal{C}^{\delta}(X_{50\ 000}))$, for δ ranging from 0 to $+\infty$, as shown in Figure 8(b). Since the convexity radius ϱ_X is infinite in this case, we have $H_1(\mathcal{C}^{\delta}(X_k)) \cong H_1(\mathcal{O}^{\delta}(X_k)) \cong H_1(X)$ for all $\delta > d_{\mathrm{H}}(X_k, X)$, and so our heuristic to determine a suitable rate $(\delta'_k)_{k\in\mathbb{N}}$ from a sample X_{k_0} for some sufficiently large $k_0 \in \mathbb{N}$ is:

1. Compute the pointwise dimension of $H_1(\mathcal{C}^{\delta}(X_{k_0}))$ as in Figure 8(b) for a range of δ;
2. Identify the transition phase of $\dim H_1(\mathcal{C}^{\delta}(X_{k_0}))$;
3. Choose a scale δ^* past this phase, so that $\delta \in [\delta^*, \infty) \mapsto \dim H_1(\mathcal{C}^{\delta}(X_{k_0}))$ is constant;
4. Define the rate $(\delta'_k)_k$ for $k \in \mathbb{N}$ as $\delta'_k := \frac{\delta^*}{\delta_{k_0}}\delta_k$.

In Figure 9, we observe that both rates δ'_k and $\hat{\delta}_k$ achieve good performance asymptotically, with a convergence rate that is at least of the order of $\left(\frac{\log(k)}{k}\right)^{1/2}$. However, the asymptotic regime is reached earlier with δ'_k than with $\hat{\delta}_k$ – while, as we said, it is not reached with δ_k.



(a) Convergence rate with $\delta'_k = \frac{1}{15}\delta_k$. **(b)** Convergence rate with $\hat{\delta}_k$.

Figure 9 Interleaving distance between the estimators and the target w.r.t. the sample size k.

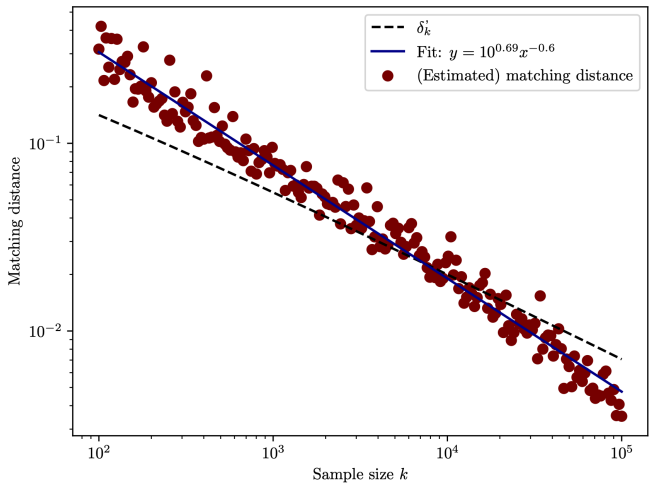


Figure 10 Estimated matching distance between the estimator $H_0(\mathcal{R}^{\delta'_k \to 2\delta'_k}(f|_{X_k}))$ and the ground truth $H_0(f)$ as a function of the sample size k under random sampling.

5.2.2 Estimating the combined functions (f_1, f_2, f_3)

We now study the function $f = (f_1, f_2, f_3)\colon [0,1]^2 \to \mathbb{R}^3$ and approximate the corresponding 3-parameter persistence module $H_0(f)$. For this we employ the estimator $H_0(\mathcal{R}^{\delta'_k \to 2\delta'_k}(f|_{X_k})) \cong H_0(\mathcal{R}^{2\delta'_k}(f|_{X_k}))$, where, for each sample size k, the scale parameter is given by the same δ'_k as in Section 5.2.1. Since computing the interleaving distance in this context is NP-hard [3], we use the matching distance [15] as a proxy, estimated via Monte Carlo with $1,000$ random lines in $\mathbb{R}^3$. The results are shown in Figure 10. Notably, the observed convergence rate matches that of the previous experiment in Figure 9(a), which suggests that the empirical convergence rate is independent of the number of parameters.

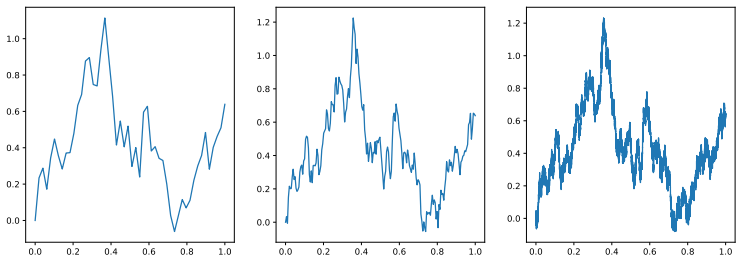


Figure 11 Samplings of a Brownian motion on $[0,1]$ with, respectively, 50, 200, and 10^6 points.

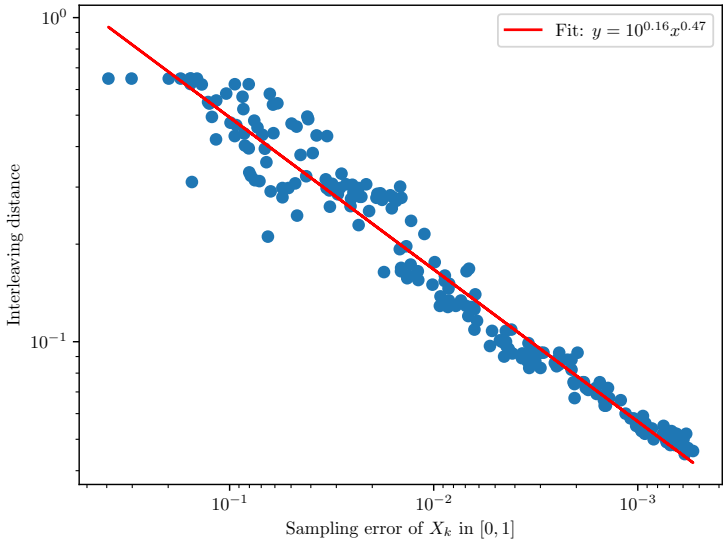


Figure 12 Log–log plot of the convergence of $H_0(\mathcal{C}^{\varepsilon_k}(f|_{X_k}))$ toward $H_0(f)$ for a Brownian motion f. The x-axis represents the sampling error $\varepsilon_k = d_{\mathrm{H}}(X_k, [0,1])$, and the y-axis represents the interleaving distance between $H_0(\mathcal{C}^{\varepsilon_k}(f|_{X_k}))$ and $H_0(f)$.

5.3 Brownian motion in 1-d (involving a Hölder function)

The Brownian motion is not Lipschitz continuous but almost surely almost everywhere ω-continuous for any modulus of continuity ω of the form $\omega\colon x \mapsto c\,x^{\frac{1}{2}-\varepsilon}$, with $\varepsilon > 0$ and $c > 0$ large enough [20, Corollaries 1.20 and 5.3]. Let $f : [0,1] \to \mathbb{R}$ be a realization of a standard Brownian motion, which can be approximated by a Rademacher random walk.

For each $k \in \mathbb{N}$, we build a k-sample X_k in $[0,1]$ according to the uniform distribution, as shown in Figure 11 for several values of k. Then, since our goal here is to illustrate Theorem 3, we assume we have access to the sampling error $\varepsilon_k := d_{\mathrm{H}}(X_k, [0,1])$ and so we use the estimator $H_0(\mathcal{C}^{\varepsilon_k}(f|_{X_k}))$ to approximate the target $H_0(f)$ for various sample sizes k. The corresponding results are shown in Figure 12. A regression analysis of the experimental data indicates that $d_{\mathrm{i}}^{\mathbf{1}}\left(H_0(\mathcal{C}^{\varepsilon_k}(f|_{X_k})), H_0(f)\right) \approx 10^{0.16}\,{\varepsilon_k}^{0.47}$, which aligns closely with the theoretical prediction from Theorem 3, stating that the interleaving distance between $H_0(\mathcal{C}^{\varepsilon_k}(f|_{X_k}))$ and $H_0(f)$ should be at most $\omega(\varepsilon_k) \lesssim {\varepsilon_k}^{1/2}$.

5.4 Timings

In the following table we report the running times for some of our experiments. Only the first one involves our algorithm for computing a presentation of $H_*(\mathcal{R}^{\bullet\to 2\bullet}(\mathscr{f}|_P))$. For all computations we used an Intel® Core™ i9-12900K CPU (5.2 GHz max) equipped with 125 GB of RAM.

Experiment	Estimator	#Params	#Points	Filt. size	Filt. time	Invariant time
2 circles	$H_1(\mathcal{R}^{\bullet\to 2\bullet}(\mathscr{f}\|_P))$	2	220	13,110	11.8 ms	6.5 ms (presentation)
Immune cells (single $\mathscr{f}_i$)	$H_1(\mathcal{C}^{\bullet}(\mathscr{f}_i\|_{X_k}))$	2	6172	24,370	864 ms	50.3 ms (Hilbert measure)
Brownian motion	$H_0(\mathcal{C}^{\varepsilon_k}(\mathscr{f}\|_{X_k}))$	1	10,000	37,359	1.21 s	5.4 ms (barcode)

References

1 Ethan André, Jingyi Li, David Loiseaux, and Steve Oudot. Estimating the persistent homology of $\mathbb{R}^n$-valued functions using function-geometric multifiltrations, 2025. arXiv:2412.04162.

2 Ulrich Bauer and Michael Lesnick. Induced matchings and the algebraic stability of persistence barcodes. *Journal of Computational Geometry*, 6, 2015. doi:10.20382/JOCG.V6I2A9.

3 Håvard Bakke Bjerkevik, Magnus Bakke Botnan, and Michael Kerber. Computing the interleaving distance is np-hard. *Foundations of Computational Mathematics*, 20(5):1237–1271, 2020. doi:10.1007/S10208-019-09442-Y.

4 Omer Bobrowski, Sayah Mukherjee, and Jonathan E. Taylor. Topological consistency via kernel estimation. *Bernoulli*, 23(1):288–328, 2017.

5 Magnus Bakke Botnan, Vadim Lebovici, and Steve Oudot. On rectangle-decomposable 2-parameter persistence modules. *Discrete & Computational Geometry*, pages 1–24, 2022.

6 Mathieu Carriere, Bertrand Michel, and Steve Oudot. Statistical analysis and parameter selection for mapper. *Journal of Machine Learning Research*, 19(12):1–39, 2018. URL: https://jmlr.org/papers/v19/17-291.html.

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

8 Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J Guibas, and Steve Y Oudot. Proximity of persistence modules and their diagrams. In *Proceedings of the twenty-fifth annual symposium on Computational geometry*, pages 237–246, 2009. doi:10.1145/1542362.1542407.

9 Frédéric Chazal, Vin De Silva, Marc Glisse, and Steve Oudot. *The structure and stability of persistence modules*, volume 10. Springer, 2016. doi:10.1007/978-3-319-42545-0.

10 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. doi:10.1007/S00454-011-9360-X.

11 Frédéric Chazal, Leonidas J Guibas, Steve Y Oudot, and Primoz Skraba. Persistence-based clustering in Riemannian manifolds. *Journal of the ACM (JACM)*, 60(6):1–38, 2013. doi:10.1145/2535927.

12 Tamal K Dey, Jan Jendrysiak, and Michael Kerber. Decomposing multiparameter persistence modules. *arXiv preprint*, 2025. arXiv:2504.08119.

13 Allen Hatcher. *Algebraic Topology*. Cambridge Univ. Press, 2001. URL: http://www.math.cornell.edu/~hatcher/.

14 Rafal Komendarczyk, Sushovan Majhi, and Will Tran. Topological stability and latschev-type reconstruction theorems for spaces of curvature bounded above. *arXiv preprint*, 2024. arXiv:2406.04259.

15 Claudia Landi. The rank invariant stability via interleavings. In *Research in computational topology*, pages 1–10. Springer, 2018.

16 Janko Latschev. Vietoris-Rips complexes of metric spaces near a closed Riemannian manifold. *Archiv der Mathematik*, 77(6):522–528, 2001.

17 David Loiseaux, Mathieu Carrière, and Andrew J. Blumberg. Multi-parameter Module Approximation: An efficient and interpretable invariant for multi-parameter persistence modules with guarantees. *Journal of Applied and Computational Topology*, 9(4):26, December 2025. doi:10.1007/s41468-025-00222-y.

18 David Loiseaux, Luis Scoccola, Mathieu Carrière, Magnus Bakke Botnan, and Steve Oudot. Stable Vectorization of Multiparameter Persistent Homology using Signed Barcodes as Measures. *Advances in Neural Information Processing Systems*, 36:68316–68342, December 2023.

19 Saunders Mac Lane. *Categories for the working mathematician*, volume 5. Springer Science & Business Media, 2013.

20 Peter Mörters and Yuval Peres. *Brownian Motion*, volume 30. Cambridge University Press, 2010.

21 Steve Oudot and Luis Scoccola. On the stability of multigraded Betti numbers and Hilbert functions. *SIAM Journal on Applied Algebra and Geometry*, 8(1):54–88, 2024. doi:10.1137/22M1489150.

22 Steve Oudot and Lukas Waas. A persistent version of Latschev's theorem. In *42nd International Symposium on Computational Geometry (SoCG 2026)*, Leibniz International Proceedings in Informatics (LIPIcs), pages 82:1–82:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2026. doi:10.4230/LIPIcs.SoCG.2026.82.

23 Alexander Rolle and Luis Scoccola. Stable and consistent density-based clustering via multiparameter persistence. *Journal of Machine Learning Research*, 25(258):1–74, 2024.

24 Oliver Vipond, Joshua A. Bull, Philip S. Macklin, Ulrike Tillmann, Christopher W. Pugh, Helen M. Byrne, and Heather A. Harrington. Multiparameter persistent homology landscapes identify immune cell spatial patterns in tumors. *Proceedings of the National Academy of Sciences*, 118(41):e2102166118, October 2021. doi:10/gpd543.