Journal of Applied and Computational Topology (2022) 6:331–369
https://doi.org/10.1007/s41468-022-00090-w

Statistical analysis of Mapper for stochastic and multivariate filters

Mathieu Carrière[1] · **Bertrand Michel**[2]

Received: 15 January 2021 / Revised: 13 December 2021 / Accepted: 7 February 2022 /
Published online: 29 March 2022

Abstract

Reeb spaces, as well as their discretized versions called Mappers, are common descriptors used in topological data analysis, with plenty of applications in various fields of science, such as computational biology and data visualization, among others. The stability and quantification of the rate of convergence of the Mapper to the Reeb space has been studied a lot in recent works (Brown et al. in CoRR. arXiv:1909.03488, 2019; Carrière and Oudot in Found Comput Math 18(6):1333–1396, 2017; Carrière et al. in J Mach Learn Res 19(12):1–39, 2018; Munch and Wang in: 32nd international symposium on computational geometry (SoCG 2016), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 51: 53:1–53:16, 2016), focusing on the case where a scalar-valued filter is used for the computation of Mapper. On the other hand, much less is known in the multivariate case, when the codomain of the filter is $\mathbb{R}^p$, and in the general case, when it is a general metric space $(\mathcal{Z}, d_{\mathcal{Z}})$, instead of $\mathbb{R}$. The few results that are available in this setting (Dey et al. in: 33rd international symposium on computational geometry (SoCG 2017), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 77, 36:1–36:16, 2017; Munch and Wang, 2016) can only handle continuous topological spaces and cannot be used as is for finite metric spaces representing data, such as point clouds and distance matrices. In this article, we introduce a slight modification of the usual Mapper construction and we give risk bounds for estimating the Reeb space using this estimator. Our approach applies in particular to the setting where the filter function used to compute Mapper is also estimated from data, such as the eigenfunctions of PCA. Our results are given with respect to the Gromov-Hausdorff distance, computed with specific filter-based pseudometrics for Mappers and Reeb spaces defined in Dey et al. (2017). We finally provide examples of this setting in statistics and machine

✉ Mathieu Carrière
mathieu.carriere@inria.fr

Bertrand Michel
bertrand.michel@ec-nantes.fr

1 DataShape, Inria Sophia Antipolis, Biot, France

2 Laboratoire de Mathématiques Jean Leray, UMR_C 6629, Ecole Centrale de Nantes, Nantes, France

learning for different kinds of target filters, as well as numerical experiments that demonstrate the relevance of our approach.

Keywords Topological data analysis · Mapper · Confidence regions

Mathematics Subject Classification 55N31 · 62R40

1 Introduction

The *Reeb space* and the *Mapper* are common descriptors of topological data analysis (see for instance Chazal and Michel 2017), that can summarize and encode the topological features of a given data set using a continuous function, often called *filter*, defined on it. As such, both objects have been used tremendously in many different fields and applications of data science, including, among others, computational biology (Carrière and Rabadán 2018; Jeitziner et al. 2019; Nicolau et al. 2011; Rizvi et al. 2017), computer graphics (Ge et al. 2011; Singh et al. 2007), or machine learning (Brüel-Gabrielsson and Carlsson 2018; Naitzat et al. 2018). Mathematically speaking, the Reeb space is a quotient space and the Mapper is a simplicial complex. Both objects are representatives of the topology of the input data set, in the sense that any topological feature that is present in these objects witnesses the presence of an equivalent one in the input data. Moreover, the Mapper can be thought of as a more tractable approximation of the Reeb space, which, as a quotient space, might be difficult to describe and compute exactly. In the simpler case where the filter function is scalar-valued, the Mapper and the Reeb space actually become combinatorial graphs, which is why they are mostly used for clustering and data visualization. Actually, even when the filter is multivariate, i.e., when its domain belongs to $\mathbb{R}^p$ with $p > 1$, it is common to only compute the skeleton in dimension 1 of the Mapper, so as to make it easy to display and interpret. Even though computation is easier, restricting to scalar-valued functions can still be a dramatic simplification, since it happens quite often in practice that either multiple filters jointly characterize the data—as is the case, for instance, of multiple driver genes explaining a disease or cell differentiation—or that the filter actually takes values in spaces more complicated than Euclidean space, such as, e.g., the space of probability distributions—as is the case where filter functions are stochastic—making Mappers and Reeb spaces computed with mere realizations of the filter (in Euclidean space) extremely limited.

In recent works, different notions of stability and convergence of the Mapper to the Reeb space, in the case where the filter function is scalar-valued, have been defined and studied (Bauer et al. 2014; Brown et al. 2019; Carrière and Oudot 2017; Carrière et al. 2018; de Silva et al. 2016), under various statistical assumptions on how data is generated. The case of multivariate and more general filter functions is however much more difficult and less understood, since the singular values of the filter function, which turn out to be critical quantities to look at in the analysis, cannot be ordered easily, and as a consequence, the natural stratification of data (that could be derived from scalar-valued Morse functions for instance) does not extend. Few available results, presented in Munch and Wang (2016) and Dey et al. (2017), prove nice approximation

inequalities for continuous spaces, but unfortunately do not apply when data is given as a finite metric space, such as a point cloud or a distance matrix, since those finite metric spaces should be thought of as approximations of the underlying continuous space, and not the space itself.

Moreover, many of previously cited works only consider the case where the values of the filter function (either scalar-valued or multivariate) are known exactly on the data points. This will not be the case if the filter function is estimated from data, and thus different from the filter function used to compute the target Reeb space, as is the case for instance for PCA filters or density filters, which are abundant in Mapper applications. This also happens tremendously in statistics and machine learning, where the underlying filter is usually a predictor, that has to be estimated with standard machine learning methods. As explained in this article, another interesting example is when the interesting and underlying filter is given by the (scalar-valued) means, or the (multivariate) histograms, of some conditional probability distributions associated to each point in the data set, and that what is given at hand are merely single realizations of these distributions. Then, the usual way of computing Mappers will clearly not work, especially if these conditional probability distributions have large variances, since single realizations are not representative at all of the means, or histograms, of their associated conditional probability distributions.

Contributions The contribution of this article is two-fold:

- We propose risk bounds for the estimation of the Reeb space with a Mapper-based estimator in the general case, that is, for any type of filter whose codomain is a complete and locally compact length space (Dmitri et al. 2001). For this, we use the Gromov-Hausdorff distance computed with filter-based pseudometrics defined on Mappers and Reeb spaces (and originally introduced in Dey et al. 2017). Our results are stated in the context where the filter used to compute the Mapper is only an estimation (usually computed from a random sample of data) of the target filter used to compute the Reeb space.
- We propose some methodology for using our Mapper-based estimator. We also provide applications and numerical experiments in statistics and machine learning, as well as examples in which the standard Mapper fails at recovering the correct topology of the data, while using our Mapper-based estimator succeeds at doing so.

The plan of this article is as follows: in Sect. 2, we recall the basics of Reeb spaces, Mappers, and we introduce the pseudometrics defined on them. Then, we show risk bounds for our Mapper-based estimator in Sect. 3. Numerical experiments and applications are presented in Sect. 4. Finally, we conclude and provide future investigations in Sect. 5.

2 Background on Reeb spaces and Mappers

In this section, we recall the definitions of the Reeb spaces and Mappers (Sect. 2.1), and we introduce the Gromov-Hausdorff distance and the filter-based pseudometrics that we use to compare them (Sect. 2.2).

2.1 Reeb spaces and Mappers

Reeb spaces and Mappers are mathematical constructions that enable to simplify and visualize the various topological structures that are present in topological spaces, through the lens of a continuous function, often called *filter*.

Reeb space Given a topological space $\mathcal{X}$ and a continuous function $f : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is a topological space, the *Reeb space* of $\mathcal{X}$ is an approximation of $\mathcal{X}$ that preserves its connectivity structures. When $f : \mathcal{X} \to \mathbb{R}$ is scalar-valued, it is usually called the *Reeb graph* (Reeb 1946).

Definition 2.1 Let $\mathcal{X}, \mathcal{Z}$ be topological spaces and $f : \mathcal{X} \to \mathcal{Z}$ be a continuous function. The *Reeb space* of $\mathcal{X}$ is the quotient space

$$\mathrm{R}_f(\mathcal{X}) = \mathcal{X}/\sim_f,$$

where, for all $x, x' \in \mathcal{X}$, one has $x \sim_f x'$ if and only if $f(x) = f(x')$ and x, x' belong to the same connected component of $f^{-1}(f(x)) = f^{-1}(f(y))$.

Moreover, the Reeb space comes with a projection $\pi : \mathcal{X} \to \mathrm{R}_f(\mathcal{X})$ defined with $\pi(x) = [x]_{\sim_f}$, where $[x]_{\sim_f}$ denotes the equivalence class of x w.r.t. the relation $\sim_f$. Since f is continuous, so is π.

Approximation with Mapper However, the Reeb space is not well-defined when data is given as a finite metric space, i.e., a point cloud or a distance matrix, in which case all preimages used to compute the Reeb space are either empty or singletons. To handle this issue, the *Mapper* was introduced in Singh et al. (2007) as a tractable approximation of the Reeb space, computed as the *nerve* of a specific cover. We first provide its definition for continuous spaces.

Definition 2.2 Let $\mathcal{X}$ be a topological space and $\mathcal{U}$ be a cover of $\mathcal{X}$, that is, a family of subsets $\{U_\alpha\}_{\alpha \in A}$ of $\mathcal{X}$ such that $\mathcal{X} \subseteq \bigcup_{\alpha \in A} U_\alpha$. The *nerve* of X is the simplicial complex $\mathcal{N}(\mathcal{V})$ defined with

$$\sigma = \{U_{\alpha_1}, \ldots, U_{\alpha_k}\} \in \mathcal{N}(\mathcal{V}) \Leftrightarrow \bigcap_{i=1}^{k} U_{\alpha_i} \neq \emptyset.$$

In words, the nerve complex is the complex whose vertices are in bijection with the cover elements, and which contains a simplex as soon as the corresponding cover elements have a non-empty intersection.

Since it can be difficult to define a cover for a general topological space $\mathcal{X}$, one can use a filter function $f : \mathcal{X} \to \mathcal{Z}$ to do so, by first covering the domain of the function $\mathcal{Z}$, then pushing back (under f) the cover elements to the space $\mathcal{X}$, and finally refining that cover of $\mathcal{X}$ into its connected components (CC).

Definition 2.3 Let $\mathcal{X}, \mathcal{Z}$ be topological spaces and $f : \mathcal{X} \to \mathcal{Z}$ be a continuous function. Moreover, let $\mathcal{U}$ be a cover of $\mathrm{im}(f)$. Let $\mathcal{V}$ be the cover of $\mathcal{X}$ defined as

$\mathcal{V} = \{V \subseteq \mathcal{X} : \exists \alpha \in A \text{ s.t. } V \text{ is a CC of } f^{-1}(U_\alpha)\}$. The *Mapper* of $\mathcal{X}, f, \mathcal{U}$ is then defined as the nerve complex

$$\mathrm{M}_{f,\mathcal{U}}(\mathcal{X}) = \mathcal{N}(\mathcal{V}).$$

Parameters and extension to point cloud When data is given as a finite metric space $\mathbb{X}_n$ with n points, the connected components are usually identified with clustering, and the nerve is computed by assessing a non-empty intersection between several cover elements as soon as there exists at least one point that is shared by all these elements. In the remaining of this article, we use graph clustering. More precisely, we assume that we have a graph G built on top of our finite metric space, and for each element U of the cover $\mathcal{U}$, we use the connected components of the subgraph $G(U)$ to compute the Mapper, where $G(U)$ is defined as

$$G(U) = (V_U, E_U), \tag{1}$$

where the vertex set V_U is $\{v \in V(G) : f(v) \in U\}$ and the edge set E_U is $\{(u, v) \in E(G) : u \in V_U, v \in V_U\}$). When G is set to be the δ-neighborhood graph G_δ, this amounts to perform single-linkage clustering (Murtagh and Contreras 2012) with parameter δ, and we let

$$\mathrm{M}_{f,\mathcal{U},G_\delta}(\mathbb{X}_n) \tag{2}$$

denote the corresponding Mapper for finite metric spaces.

Moreover, when $\mathcal{Z} = \mathbb{R}^p$, it is very usual to define a cover $\mathcal{U}$ with hypercubes by covering every single dimension of $\mathbb{R}^p$ with intervals of length $r > 0$ and overlap percentage $g \in [0, 1]$, and then by taking the Euclidean products of these intervals. Note that r and g are often called the *resolution* and the *gain* of the cover respectively. We let $\mathcal{U}(r, g)$ denote this particular type of cover. Note however that this strategy becomes quickly very expensive, and thus prohibitive, when the dimension p is large. Actually, even for moderate values, e.g., $p = 10$, the computation can become very costly if the resolution is too small or the gain is too large. Moreover, from a statistical perspective, such a naive strategy requires a number of observations which increases exponentially with the dimension, due to the curse of dimensionality. It is thus essential to propose greedy methods to define efficient covers in such situations. In Sect. 3.1, we provide alternative and computationally feasible strategies to cover the filter domain using thickenings of partitions.

It has been shown in recent works (Brown et al. 2019; Carrière and Oudot 2017; Carrière et al. 2018; Munch and Wang 2016) that the Mapper actually approximates the Reeb space under various assumptions and metrics when the filter is scalar-valued. In the next section, we introduce the filter-based pseudometrics that we will use for comparing Mappers and Reeb spaces with the Gromov–Hausdorff distance.

2.2 The filter-based pseudometric

The filter-based pseudometric, introduced in Bauer et al. (2014); Dey et al. (2017), basically measures the diameter of continuous paths between any two points *relative to the filter values*. In order to define this distance properly, one needs a metric $d_{\mathcal{Z}}$ defined on the domain $\mathcal{Z}$ of the continuous function f.

Definition 2.4 Let $\mathcal{X}$ be a topological space, $(\mathcal{Z}, d_{\mathcal{Z}})$ be a metric space and $f : \mathcal{X} \to \mathcal{Z}$ be a continuous function defined on it. The *filter-based pseudometric* $d_f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined as

$$d_f(x, x') = \inf_{\gamma \in \Gamma(x,x')} \max_{t,t' \in [0,1]} d_{\mathcal{Z}}(f \circ \gamma(t), f \circ \gamma(t')) = \inf_{\gamma \in \Gamma(x,x')} \operatorname{diam}_{\mathcal{Z}}(f \circ \gamma),$$

where $\Gamma(x, x')$ denotes the set of all continuous paths $\gamma : [0, 1] \to \mathcal{X}$ such that $\gamma(0) = x$ and $\gamma(1) = x'$, and $\operatorname{diam}_{\mathcal{Z}}$ denotes the *diameter* of a subset of $\mathcal{Z}$, i.e., $\operatorname{diam}_{\mathcal{Z}}(Z) = \sup_{z,z' \in Z} d_{\mathcal{Z}}(z, z')$ for any $Z \subseteq \mathcal{Z}$.

This distance can be used for the Reeb space itself, thanks to the following lemma.

Lemma 2.5 *Let $\mathcal{X}$ be a topological space, $(\mathcal{Z}, d_{\mathcal{Z}})$ be a metric space and $f : \mathcal{X} \to \mathcal{Z}$ be a continuous function defined on it. Let $\mathrm{R}_f(\mathcal{X})$ be the corresponding Reeb space with associated projection $\pi : \mathcal{X} \to \mathrm{R}_f(\mathcal{X})$, and $r, r' \in \mathrm{R}_f(\mathcal{X})$. Finally, let $x, y \in \pi^{-1}(r)$ and $x', y' \in \pi^{-1}(r')$. Then, one has $d_f(x, x') = d_f(y, y')$.*

Proof Since $x, y \in \pi^{-1}(r)$, one has $x \sim_f y$, meaning that $f(x) = f(y)$, and x, y are in the same connected component of $f^{-1}(f(x))$. Thus, there exists a path $\gamma_x^y : [0, 1] \to \mathcal{X}$ such that $\gamma_x^y(0) = x$, $\gamma_x^y(1) = y$, and $f \circ \gamma_x^y(t) = f(x)$ for any $t \in [0, 1]$. Similarly, there exists a path $\gamma_{y'}^{x'} : [0, 1] \to \mathcal{X}$ such that $\gamma_{y'}^{x'}(0) = y'$, $\gamma_{y'}^{x'}(1) = x'$, andj $f \circ \gamma_{y'}^{x'}(t) = f(x')$ for any $t \in [0, 1]$.

Now, let $\Gamma(x, x')$ and $\Gamma(y, y')$ denote the sets of all continuous paths between x, x' and y, y' respectively. Note that, for any path $\gamma' \in \Gamma(y, y')$, one has (up to renormalization) that $\gamma := \gamma_{y'}^{x'} \circ \gamma' \circ \gamma_x^y \in \Gamma(x, x')$. Moreover, since f is constant on γ_x^y and $\gamma_{y'}^{x'}$, one has $\operatorname{diam}_{\mathcal{Z}}(f \circ \gamma) = \operatorname{diam}_{\mathcal{Z}}(f \circ \gamma')$. Then, $d_f(x, y) \leq \operatorname{diam}_{\mathcal{Z}}(f \circ \gamma) = \operatorname{diam}_{\mathcal{Z}}(f \circ \gamma')$. Since this is true for any $\gamma' \in \Gamma(y, y')$, one has $d_f(x, x') \leq d_f(y, y')$.

Using the same arguments, one also has $d_f(y, y') \leq d_f(x, x')$ by symmetry. □

Lemma 2.5 allows to define the following extension of the pseudometric to Reeb spaces.

Definition 2.6 Let $\mathcal{X}$ be a topological space, $(\mathcal{Z}, d_{\mathcal{Z}})$ be a metric space and $f : \mathcal{X} \to \mathcal{Z}$ be a continuous function defined on it. Let d_f be the corresponding pseudometric, and $\mathrm{R}_f(\mathcal{X})$ be the corresponding Reeb space. Then, the Reeb space $\mathrm{R}_f(\mathcal{X})$ can also be equipped with a pseudometric $\tilde{d}_f$ using the projection π. For any two equivalence classes $r, r' \in \mathrm{R}_f(\mathcal{X})$,

$$\tilde{d}_f(r, r') = d_f(x, x') \text{ for arbitrary } x \in \pi^{-1}(r) \text{ and } x' \in \pi^{-1}(r').$$

Finally, we also define a pseudometric between the nodes of the Mapper $\mathrm{M}_{f,\mathcal{U}}(\mathcal{X})$, for a given cover $\mathcal{U}$ of $\mathrm{im}(f)$. Recall that the nodes of the Mapper are the vertices of the cover $\mathcal{N}(\mathcal{V})$ (see Definition 2.3), and hence each node v corresponds to a connected component of $f^{-1}(U)$ for some $U \in \mathcal{U}$. Thus, we can associate an arbitrary (but distinct) element $z_v \in U$ for each v. Let $f_{\mathcal{U}}$ be the corresponding map:

$$f_{\mathcal{U}} : \begin{cases} V(\mathrm{M}_{f,\mathcal{U}}(\mathcal{X})) \to \mathcal{Z} \\ v \qquad\qquad\quad \mapsto z_v \end{cases} \tag{3}$$

Definition 2.7 Let $\mathcal{X}$ be a topological space, $(\mathcal{Z}, d_{\mathcal{Z}})$ be a metric space and $f : \mathcal{X} \to \mathcal{Z}$ be a continuous function defined on it. Let $\mathcal{U}$ be a cover of $\mathrm{im}(f)$, and $f_{\mathcal{U}}$ be an associated map as per Equ. (3). The filter-based pseudometric is then defined with

$$\tilde{d}_{f,\mathcal{U}}(v, v') = \inf_{\gamma \in \Gamma(v,v')} \max_{p,q \in \gamma} d_{\mathcal{Z}}(f_{\mathcal{U}}(p), f_{\mathcal{U}}(q)),$$

where $\Gamma(v, v')$ denotes the set of all paths between v and v' in $\mathrm{M}_{f,\mathcal{U}}(\mathcal{X})$, that is, $\Gamma(v, v')$ is of the form

$$\Gamma(v, v') = \{p_1, \ldots, p_n : n \in \mathbb{N}, p_1 = v, p_n = v', (p_i, p_{i+1}) \\ \text{is an edge of } \mathrm{M}_{f,\mathcal{U}}(\mathcal{X}), \forall 1 \leq i \leq n-1\}.$$

One of the main benefits of defining pseudometrics for $\mathcal{X}$, its Reeb space $\mathrm{R}_f(\mathcal{X})$ and its Mapper $\mathrm{M}_{f,\mathcal{U}}(\mathcal{X})$, is that we can then compare these spaces using the *Gromov-Hausdorff distance*.

Definition 2.8 [*Dmitri et al.* 2001, *Theorem 7.3.25*] Let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ be metric spaces. A *correspondence* C between $\mathcal{X}$ and $\mathcal{Y}$ is a subset of $\mathcal{X} \times \mathcal{Y}$ such that $\pi_1 : C \to \mathcal{X}$ defined with $\pi_1 : (x, y) \mapsto x$ and $\pi_2 : C \to \mathcal{Y}$ defined with $\pi_2 : (x, y) \mapsto y$ are surjective. The *distortion* of a correspondence C is then defined as $\mathrm{dis}(C) = \sup\{|d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y')| : (x, y), (x', y') \in C\}$. Finally, the *Gromov-Hausdorff distance* between $\mathcal{X}$ and $\mathcal{Y}$ is defined as

$$d_{\mathrm{GH}}((\mathcal{X}, d_{\mathcal{X}}), (\mathcal{Y}, d_{\mathcal{Y}})) = \frac{1}{2}\inf_C \mathrm{dis}(C),$$

where C ranges over all correspondences between $\mathcal{X}$ and $\mathcal{Y}$.

Note that if $d_{\mathcal{X}}$ or $d_{\mathcal{Y}}$ is only a pseudometric, then the Gromov-Hausdorff distance is still well-defined, but becomes a pseudometric too. Now, according to the following proposition, the spaces $\mathcal{X}$ and $\mathrm{R}_f(\mathcal{X})$ (equipped with their pseudometrics) are actually the same when compared with the Gromov-Hausdorff pseudometric.

Proposition 2.9 *Let $\mathcal{X}$ be a topological space, $(\mathcal{Z}, d_{\mathcal{Z}})$ be a metric space and $f : \mathcal{X} \to \mathcal{Z}$ be a continuous function defined on it. Then*

$$d_{\mathrm{GH}}((\mathcal{X}, d_f), (\mathrm{R}_f(\mathcal{X}), \tilde{d}_f)) = 0.$$

Proof Since π is surjective, we can use it to define a correspondence C between $\mathcal{X}$ and $\mathrm{R}_f(\mathcal{X})$, with $C = \{(x, \pi(x)) : x \in \mathcal{X})\}$. Then, $d_{\mathrm{GH}}((\mathcal{X}, d_f), (\mathrm{R}_f(\mathcal{X}), \tilde{d}_f)) \leq \frac{1}{2}\mathrm{dis}(C) = \frac{1}{2}\sup\{|d_f(x, x') - \tilde{d}_f(\pi(x), \pi(x'))| : x, x' \in \mathcal{X}\} = 0$ by definition of $\tilde{d}_f$. □

Similar bounds can be obtained for the Mapper as well. These bounds depend on a key quantity, called the cover *resolution*, that we define below.

Definition 2.10 Let $\mathcal{X}$ be a topological space, $(\mathcal{Z}, d_{\mathcal{Z}})$ be a metric space and $f : \mathcal{X} \to \mathcal{Z}$ be a continuous function defined on it. Let $\mathcal{U}$ be a cover of $\mathrm{im}(f)$. The *resolution of* $\mathcal{U}$ w.r.t. *the filter* f, denoted by $\mathrm{res}(\mathcal{U}, f)$, is defined as $\mathrm{res}(\mathcal{U}, f) := \max_{\alpha \in A} \sup_{u,v \in U_\alpha \cap \mathrm{im}(f)} d_{\mathcal{Z}}(u, v) = \max_{\alpha \in A} \mathrm{diam}_{\mathcal{Z}}(U_\alpha \cap \mathrm{im}(f))$.

It turns out that Mappers and Reeb spaces equipped with their respective pseudo-metrics are actually close for covers with small resolutions.

Theorem 2.11 [Dey et al. (2017), Theorem 32] *Let $\mathcal{X}$ be a topological space, $(\mathcal{Z}, d_{\mathcal{Z}})$ be a metric space and $f : \mathcal{X} \to \mathcal{Z}$ be a continuous function defined on it. Let $\mathcal{U}$ be a cover of* $\mathrm{im}(f)$. *Then*

$$d_{\mathrm{GH}}((\mathrm{M}_{f,\mathcal{U}}(\mathcal{X}), \tilde{d}_{f,\mathcal{U}}), (\mathrm{R}_f(\mathcal{X}), \tilde{d}_f)) \leq 5 \cdot \mathrm{res}(\mathcal{U}, f).$$

Cover resolutions are also useful because they allow one to control the variations induced by the arbitrary choices of Eq. (3). In particular, the smaller the resolution, the less difference those arbitrary choices make in terms of the Gromov-Hausdorff pseudometric.

Proposition 2.12 *Let $\mathcal{X}$ be a topological space, $(\mathcal{Z}, d_{\mathcal{Z}})$ be a metric space and $f : \mathcal{X} \to \mathcal{Z}$ be a continuous function defined on it. Let $\mathcal{U}$ be a cover of* $\mathrm{im}(f)$, *and $f_{\mathcal{U}}$, $f'_{\mathcal{U}}$ be two associated maps as per Eq. (3), with corresponding pseudometrics $\tilde{d}_{f,\mathcal{U}}$ and $\tilde{d}'_{f,\mathcal{U}}$ respectively. Then, one has*

$$d_{\mathrm{GH}}((\mathrm{M}_{f,\mathcal{U}}(\mathcal{X}), \tilde{d}_{f,\mathcal{U}}), (\mathrm{M}_{f,\mathcal{U}}(\mathcal{X}), \tilde{d}'_{f,\mathcal{U}})) \leq \mathrm{res}(\mathcal{U}, f)$$

Proof Let C be the trivial correspondence $C = \{(v, v) : v \in \mathrm{M}_{f,\mathcal{U}}(\mathcal{X})\}$. Then, one has the following inequality: $d_{\mathrm{GH}}((\mathrm{M}_{f,\mathcal{U}}(\mathcal{X}), \tilde{d}_{f,\mathcal{U}}), (\mathrm{M}_{f,\mathcal{U}}(\mathcal{X}), \tilde{d}'_{f,\mathcal{U}})) \leq \frac{1}{2}\mathrm{dis}(C) = \frac{1}{2}\sup\{|\tilde{d}_{f,\mathcal{U}}(v, v') - \tilde{d}'_{f,\mathcal{U}}(v, v')|, : v, v' \in \mathrm{M}_{f,\mathcal{U}}(\mathcal{X})\}$. Now, let $v, v' \in \mathrm{M}_{f,\mathcal{U}}(\mathcal{X})$, and let $\Gamma(v, v')$ be the set of all paths between v and v' in $\mathrm{M}_{f,\mathcal{U}}(\mathcal{X})$. Let $\gamma \in \Gamma(v, v')$. Then,

$$\begin{aligned}
\tilde{d}_{f,\mathcal{U}}(v, v') &\leq \mathrm{diam}_{\mathcal{Z}}(f_{\mathcal{U}} \circ \gamma) \\
&= d_{\mathcal{Z}}(f_{\mathcal{U}}(p), f_{\mathcal{U}}(q)) \text{ for some } p, q \in \gamma \\
&\leq d_{\mathcal{Z}}(f_{\mathcal{U}}(p), f'_{\mathcal{U}}(p)) + d_{\mathcal{Z}}(f'_{\mathcal{U}}(p), f'_{\mathcal{U}}(q)) + d_{\mathcal{Z}}(f'_{\mathcal{U}}(q), f_{\mathcal{U}}(q)) \\
&\leq \mathrm{res}(\mathcal{U}, f) + \mathrm{diam}_{\mathcal{Z}}(f'_{\mathcal{U}} \circ \gamma) + \mathrm{res}(\mathcal{U}, f).
\end{aligned}$$

Since this is true for any $\gamma \in \Gamma(v, v')$, one has $\tilde{d}_{f,\mathcal{U}}(v, v') \leq \tilde{d}'_{f,\mathcal{U}}(v, v') + 2\text{res}(\mathcal{U}, f)$. Using the same arguments, one also has $\tilde{d}'_{f,\mathcal{U}}(v, v') \leq \tilde{d}_{f,\mathcal{U}}(v, v') + 2\text{res}(\mathcal{U}, f)$ by symmetry. Hence, for any $v, v' \in \mathrm{M}_{f,\mathcal{U}}(\mathcal{X})$, one has $|\tilde{d}_{f,\mathcal{U}}(v, v') - \tilde{d}'_{f,\mathcal{U}}(v, v')| \leq 2\text{res}(\mathcal{U}, f)$, which leads to the result. □

3 Reeb space inference

In this section, we propose a new Mapper-based estimator for Reeb spaces. Its definition requires a few more assumptions on the domain $\mathcal{Z}$ and codomain $\mathcal{X}$, namely they are assumed to be *length spaces* [Dmitri et al. (2001), Definition 2.1.6] that are also complete and locally compact.

Definition 3.1 Let $\mathcal{X}$ be a topological space. A *length structure* on $\mathcal{X}$ is a family of paths $\mathcal{P}$ together with a length function $L : \mathcal{P} \to \mathbb{R}_+ \cup \{+\infty\}$ that satisfy the following assumptions[1]:

- $\mathcal{P}$ is closed under restrictions and concatenation of paths,
- L is additive, continuous and consistent with the topology of $\mathcal{X}$.

A length structure comes with an associated *intrinsic metric* d_L on X, defined with

$$d_L(x, x') = \inf \{L(\gamma) \,:\, \gamma \in \mathcal{P}, \gamma(0) = x, \gamma(1) = x'\}.$$

A metric $d_{\mathcal{X}}$ on $\mathcal{X}$ is called *intrinsic* if it is the intrinsic metric of a length structure on $\mathcal{X}$. A metric space $(\mathcal{X}, d_{\mathcal{X}})$ is called a *length space* if its metric $d_{\mathcal{X}}$ is intrinsic.

From now on, both $\mathcal{X}$ and $\mathcal{Z}$ are assumed to be complete and locally compact length spaces. We will use $|\cdot|$ to denote the corresponding length functions. A key property of length spaces that are complete and locally compact that we will use often in this article, is that *shortest paths* exist for every pair of points in $\mathcal{X}$ and $\mathcal{Z}$, which we define below.

Definition 3.2 Let $(\mathcal{X}, d_{\mathcal{X}})$ be a complete and locally compact length space. Let $x, x' \in \mathcal{X}$. A *shortest path* between x and x' is a continuous path $\gamma^* : [0, 1] \to \mathcal{X}$ with $\gamma^*(0) = x$, $\gamma^*(1) = x'$, such that, for any other continuous path $\gamma : [0, 1] \to \mathcal{X}$ with $\gamma(0) = x$, $\gamma(1) = x'$, one has $|\gamma^*| \leq |\gamma|$ and $|\gamma^*| = d_{\mathcal{X}}(x, x')$.

Proposition 3.3 [Dmitri et al. (2001), Theorem 2.5.23] *Let $(\mathcal{X}, d_{\mathcal{X}})$ be a complete and locally compact length space. Then there exists a shortest path for any given pair of points in $\mathcal{X}$.*

The main idea behind our estimator is to first compute a refinement of the input point cloud in order to remove its pathological elements (w.r.t. the cover used for computing the estimator). These elements are the so-called *element-crossing edges*, defined in

[1] The interested reader can find formal statements associated to these assumptions in [Dmitri et al. (2001), Section 2.1.1].

Sect. 3.2. Then, our estimator is defined as the standard Mapper estimator for this refined point cloud. We first introduce a raw version of the estimator without calibrating the parameters. This calibration is then detailed further and allows to provide a risk bound for our corresponding estimator.

3.1 A Mapper-based estimator

In this section, we introduce our Mapper based estimator in a deterministic setting. Assume that two point clouds $\mathbb{X}_n$ and $\mathbb{Z}_n$ are given: $\mathbb{X}_n = (x_1, \dots, x_n)$ and $\mathbb{Z}_n = (z_1, \dots, z_n)$ such that for any i, one has $(x_i, z_i) \in \mathcal{X} \times \mathcal{Z}$ and $z_i = \hat{f}(x_i)$. The function $\hat{f} : \mathbb{X}_n \to \mathcal{Z}$ is an approximation of a "true" and unknown filter function $f : \mathcal{X} \to \mathcal{Z}$. In some settings, the true exact filter f is known and then $\hat{f} = f|_{\mathbb{X}_n}$.

Point cloud and embedded graph We let G_δ be the (metric) neighborhood graph built on top of $\mathbb{X}_n$ with parameter δ, that is, any pair $\{x_i, x_j\} \subseteq \mathbb{X}_n$ creates an edge in G_δ, with length $d_{\mathcal{X}}(x_i, x_j)$, and parameterized with a shortest path between x_i and x_j, if and only if $d_{\mathcal{X}}(x_i, x_j) \leq \delta$. We then define the corresponding *embedded graphs* as $G_\delta^{\mathcal{Z}}$ (resp. $\hat{G}_\delta^{\mathcal{Z}}$) in $\mathcal{Z}$, with vertices $\{f(x_i) : x_i \in \mathbb{X}_n\}$ (resp. $\{\hat{f}(x_i) : x_i \in \mathbb{X}_n\}$) and whose edges are geometric realizations of edges of G_δ, that is, shortest paths in $\mathcal{Z}$[2]. When $\mathcal{Z}$ is a normed vector space (such as a Banach space), this corresponds to linear interpolations in $\mathcal{Z}$. We finally extend $f|_{\mathbb{X}_n}$ (resp. $\hat{f}$) to a function $f_{\mathrm{SP}} : G_\delta \to G_\delta^{\mathcal{Z}}$ (resp. $\hat{f}_{\mathrm{SP}} : G_\delta \to \hat{G}_\delta^{\mathcal{Z}}$), which maps the interiors of the edges of G_δ to the corresponding interiors of the shortest paths in $\mathcal{Z}$ (where SP stands for Shortest Paths). Note that $f|_{\mathbb{X}_n}$ and f_{SP} (resp. $\hat{f}$ and $\hat{f}_{\mathrm{SP}}$) coincide on $\mathbb{X}_n$, so we will only use f_{SP} (resp. $\hat{f}_{\mathrm{SP}}$) when applied to (interiors of) edges of G_δ.

Graph refinement Our Mapper-based estimator is defined as the standard Mapper computed on a refinement of the graph G_δ. For $s \in \mathbb{N}^* := \mathbb{N}\backslash\{0\}$, we subdivide each edge of G_δ with s points. Let $G_{\delta,s}$ be the resulting graph (on which f_{SP} and $\hat{f}_{\mathrm{SP}}$ are still well-defined), $\mathbb{X}_{n,s}$ be the refined point cloud, and $G_{\delta,s}^{\mathcal{Z}}$, $\hat{G}_{\delta,s}^{\mathcal{Z}}$ be the refined embedded graphs.

Cover Let $\mathcal{U}$ be a finite cover of $\mathrm{im}(\hat{f}_{\mathrm{SP}})$, which can be data dependent. For now, we assume that this cover is given. We will discuss the construction of $\mathcal{U}$ further in this article.

Estimator We are now in position to define our Mapper based estimator. It is defined with (2) as:

$$\mathrm{M}_n := \mathrm{M}_{\hat{f}_{\mathrm{SP}}, \mathcal{U}, G_{\delta,s}}(\mathbb{X}_{n,s}), \tag{4}$$

and it can be equipped with the pseudometric $\tilde{d}_{\hat{f}_{\mathrm{SP}}, \mathcal{U}}$ of Definition 2.7. See Fig. 1 for an illustration of the construction of M_n. For defining the estimator, we need to choose the scale parameters s and δ, and we discuss this question further in the next section.

[2] Our construction actually works for arbitrary paths. However, some quantities that are necessary for computing the estimator (such as ℓ) might be easier to compute when working with shortest paths, so we stick to those paths in this article.

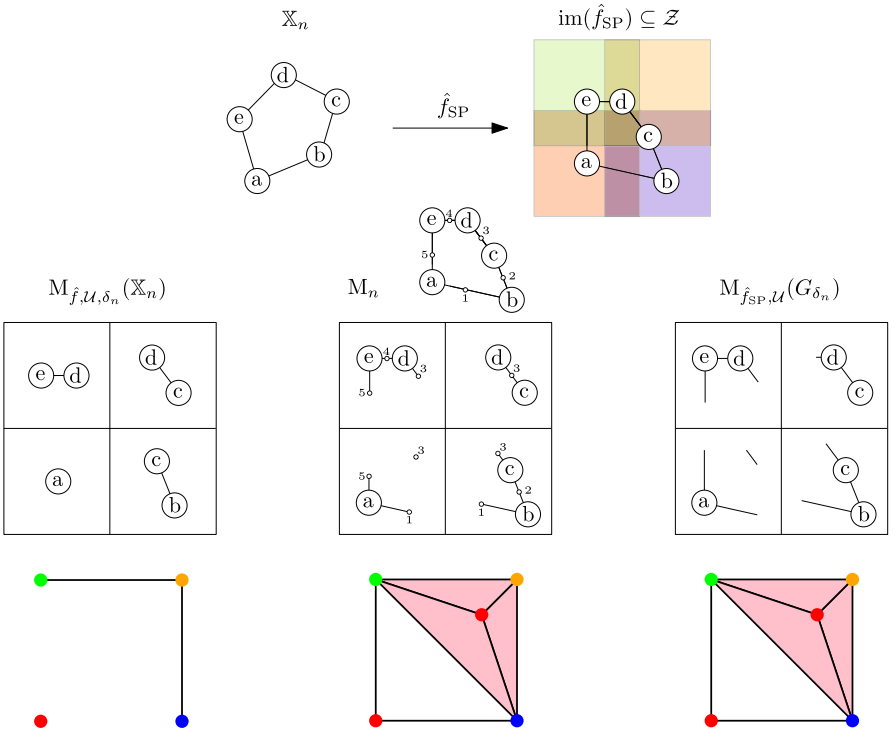

Fig. 1 Example of our estimator M_n on a dataset of 5 points. Upper left: dataset $\mathbb{X}_n$ with $n = 5$ points a, b, c, d, e. The edges between the points are computed with a neighborhood graph with parameter δ_n. Upper right: cover of $\mathrm{im}(\hat{f}_{\mathrm{SP}})$ with four squares. The edges between the points in $\mathcal{Z}$ are shortest paths in $\mathcal{Z}$. Lower left: preimages of the four squares for the standard Mapper on the point cloud and corresponding simplicial complex computed with hierarchical clustering with parameter δ_n. Lower right: preimages of the four squares for the standard Mapper on the metric neighborhood graph. Lower middle: Our estimator is computed by refining the neighborhood graph (with five extra nodes 1, 2, 3, 4, 5) and by using this new graph to compute the connected components and the intersections between them. As one can see, our estimator is an easily computable, combinatorial object like the usual Mapper, and is able to recover the correct topology while the standard Mapper does not

3.2 Risk bound and parameter calibration

We now give our main result in the *Stochastic Filter setting*, i.e., when the data points $X_1, \ldots, X_n$ are sampled i.i.d. from a distribution P and when the function $\hat{f} : \mathbb{X}_n \to \mathcal{Z}$ is allowed to be data dependent. In this setting, the Z_i's are thus also i.i.d. random variables. Our result is based on the following geometric and probabilistic assumptions:

- **(H1) Support assumption** The support $\mathcal{X}$ of P is a compact submanifold $\mathcal{X} \subseteq \mathbb{R}^D$ with positive reach $\mathrm{rch}(\mathcal{X})$ (see Appendix A).

The neighborhood graph G_δ is built with Euclidean norm $\|\cdot\|$ in $\mathbb{R}^D$. Let $D_{\mathcal{X}} < \infty$ denote the diameter of $\mathcal{X}$ (in the Euclidean distance): $D_{\mathcal{X}} = \sup\{\|x-y\| : x, y \in \mathcal{X}\}$. Hereafter, we call $\mathrm{rch}(\mathcal{X})$ and $D_{\mathcal{X}}$ the *geometric parameters of* $\mathcal{X}$.

- **(H2) Measure assumption** The probability measure P is (a, b)-standard, i.e., $P(B(x, r)) \geq \min\{1, ar^b\}$, for all $x \in \mathcal{X}$ and $r > 0$, where $B(x, r) = \{y \in \mathbb{R}^D : \|y - x\| \leq r\}$.

We now characterize the regularity of f with a *modulus of continuity for* f, i.e., a function ω defined on $\mathbb{R}^+$ which satisfies

1. $\omega(\delta) \to \omega(0) = 0$ as $\delta \to 0$;
2. ω is non negative and non-decreasing on $\mathbb{R}^+$;
3. ω is subadditive : $\omega(\delta_1 + \delta_2) \leq \omega(\delta_1) + \omega(\delta_2)$ for any $\delta_1, \delta_2 > 0$;
4. ω is continuous on $\mathbb{R}^+$.

and such that

$$|f(x) - f(x')| \leq \omega(\|x - x'\|),$$

for any $x, x' \in \mathcal{X}$.

- **(H3) Filter regularity assumption** The true filter $f : \mathcal{X} \to \mathcal{Z}$ is a continuous function on $\mathcal{X}$ which admits a modulus of continuity ω such that $x \in \mathbb{R}^+ \mapsto \frac{\omega(x)}{x}$ is a non-increasing function on $\mathbb{R}^+$.

Finally, we will make the following assumption on the cover:

- **(H4) Cover assumption** The cover $\mathcal{U}$ is assumed to cover $\mathrm{im}(\hat{f}_{\mathrm{SP}})$.

For calibrating the estimator parameters, we need to introduce the notion of *element-crossing edges*. Such edges are pathological in the sense that they may prevent to recover the correct topology of the underlying Reeb space. Given a simplex $\sigma = \{U_{\alpha_1}, \ldots, U_{\alpha_p}\}$ in the nerve $\mathcal{N}(\mathcal{U})$, we let $U_\sigma = \cap_{i=1}^p U_{\alpha_i}$.

Definition 3.4 Let $X_i, X_j \in \mathbb{X}_n$ such that the edge $e = (X_i, X_j)$ belongs to G_δ. Let $\hat{f}_{\mathrm{SP}}(e)$ be the corresponding edge in $\hat{G}_\delta^{\mathcal{Z}}$. We say that e is an *element-crossing edge with respect to the cover* $\mathcal{U}$ if there exists $\sigma \in \mathcal{N}(\mathcal{U})$ such that $\hat{f}_{\mathrm{SP}}(e) \cap U_\sigma \neq \emptyset$, $\hat{f}(X_i) \notin U_\sigma$ and $\hat{f}(X_j) \notin U_\sigma$.

In other words, e is an element-crossing edge with respect to the cover $\mathcal{U}$ if the shortest path $\hat{f}_{\mathrm{SP}}(e)$ goes through U_σ, even though its endpoints $\hat{f}(X_i)$ and $\hat{f}(X_j)$ are outside U_σ. In this case we say that U_σ is *crossed* by e. Note that element-crossing edges are generalizations of *interval* and *intersection-crossing edges*, as defined in Carrière and Oudot (2017). We then define:

$$\ell(\mathbb{X}_n, \hat{f}, \mathcal{U}) = \inf\left\{|\tilde{e}| \,:\, \tilde{e} \text{ is a CC of } \hat{f}_{\mathrm{SP}}(e) \cap U_\sigma \text{ for some } e \in G_\delta \text{ and } \sigma \in \mathcal{N}(\mathcal{U}) \text{ s.t. } U_\sigma \text{ is crossed by } e\right\},$$

where $|\cdot|$ denotes the length in $\mathcal{Z}$ and CC is a shorthand for connected component. In other words, $\ell(\mathbb{X}_n, \hat{f}, \mathcal{U})$ is the length of the smallest connected path in the intersection between an edge of $\hat{G}_\delta^{\mathcal{Z}}$ and a cover element or intersection, such that the edge endpoints do not belong to this cover element or intersection.

For calibrating the parameters, we also need to introduce the modulus of continuity of $\hat{f}_{\mathrm{SP}}$:

$$\hat{\omega}_{\mathrm{SP}}(u) = \sup\left\{ d_{\mathcal{Z}}(\hat{f}_{\mathrm{SP}}(x), \hat{f}_{\mathrm{SP}}(x')) \,:\, \|x - x'\| \leq u \text{ and } x, x' \text{ belong to the same edge of } G_\delta \right\},$$

where $|\cdot|$ denotes the edge length in G_δ.

We are now in position to define the calibrations for δ and s. We follow a similar strategy as in Carrière et al. (2018). Let d_{H}^E denote the Hausdorff distance computed with Euclidean distances.

- **Choice for** δ. For some arbitrary $\beta > 0$, let $t(n) = \lfloor n/(\log(n))^{1+\beta} \rfloor$. We take

$$\delta = \delta_n = d_{\mathrm{H}}^E(\tilde{\mathbb{X}}_{t(n)}, \mathbb{X}_n) = \max_{X \in \mathbb{X}_n} \min_{X' \in \tilde{\mathbb{X}}_{t(n)}} \|X - X'\|, \tag{5}$$

  where $\tilde{\mathbb{X}}_{t(n)}$ is a random subsample of size $t(n)$ drawn uniformly from $\mathbb{X}_n$ with replacement.
- **Choice for** s. Let $\ell = \ell(\mathbb{X}_n, \hat{f}, \mathcal{U})$, we take

$$s \geq s_n := \left\lfloor \frac{\delta_n}{\hat{\omega}_{\mathrm{SP}}^{-1}(\ell/2)} \right\rfloor \text{ if } \ell/2 \in \mathrm{im}(\hat{\omega}_{\mathrm{SP}}) \tag{6}$$

  and $s_n = 0$ (that is, we do not refine G_δ) otherwise. By convention, we also let $s_n = +\infty$ if $\ell = 0$, which happens with null probability.

Under the previous assumptions and with the definitions of s_n and δ_n given above, we can provide the following risk bound of our Mapper based estimator:

Theorem 3.5 *Under assumptions* (H1), (H2), (H3) *and* (H4)*, the following inequality is true:*

$$\mathbb{E}\left[d_{\mathrm{GH}}((\mathrm{M}_n, \tilde{d}_{\hat{f}_{\mathrm{SP}},\mathcal{U}}), (\mathrm{R}_f(\mathcal{X}), \tilde{d}_f)) \right] \leq 5\mathbb{E}\left[\mathrm{res}(\mathcal{U}, \hat{f}_{\mathrm{SP}}) \right]$$
$$+ C\omega\left(C' \frac{\log(n)^{(2+\beta)/b}}{n^{1/b}} \right) + 2\mathbb{E}\left[\|f_{\mathrm{SP}} - \hat{f}_{\mathrm{SP}}\|_\infty \right],$$

where the constants C, C' only depends on a, b and the geometric parameters of $\mathcal{X}$, and where the third term is defined with $\|f_{\mathrm{SP}} - \hat{f}_{\mathrm{SP}}\|_\infty := \sup_{x \in G_{\delta_n}} d_{\mathcal{Z}}(f_{\mathrm{SP}}(x), \hat{f}_{\mathrm{SP}}(x))$.

The proof is given in Appendix B. This result is very general and holds for both deterministic covers and random covers that are defined from the data in a measurable way (in which case the expectation can also be written conditionally to the cover). Note also that our bound does not depend on the z_v's (defined in Eq. (3)), and thus holds as long as the z_v's are defined in a measurable way. We discuss cover choices and corresponding upper bounds on their resolutions in Sect. 3.3. Moreover, even though the third term might be difficult to control for general length spaces, when $\mathcal{Z}$ is a Hilbert space it actually reduces to

$$\mathbb{E}\left[\|f_{\mathrm{SP}} - \hat{f}_{\mathrm{SP}}\|_\infty \right] = \mathbb{E}\left[\|(f - \hat{f})|_{\mathbb{X}_n}\|_\infty \right],$$

since shortest paths are straight lines in such spaces. This is the case for instance when $\mathcal{Z}$ is a reproducing kernel Hilbert space (RKHS), which we study in more details further in Sect. 3.4.

Parameter calibrations Theorem 3.5 relies on the calibration of the parameters of our Mapper-based estimator M_n. In particular, the choice we make for the graph refinement parameter s requires to: first, upper bound the modulus of continuity $\hat{\omega}_{\mathrm{SP}}$ of $\hat{f}_{\mathrm{SP}}$, and second, to compute the smallest connected path $\ell(\mathbb{X}_n, \hat{f}, \mathcal{U})$. Controlling $\hat{\omega}_{\mathrm{SP}}$ is not possible in general, but for standard filters such as KPCA filters (see Sect. 3.4), $\hat{f}$ and $\hat{f}_{\mathrm{SP}}$ are Lipschitz functions and hence $\hat{\omega}_{\mathrm{SP}}$ can be easily bounded by the corresponding Lipschitz constant. Next, computing—or at least lower bounding—the quantity $\ell(\mathbb{X}_n, \hat{f}, \mathcal{U})$ is difficult for a general cover $\mathcal{U}$. However, it can be done exactly for particular covers, such as the ones induced by thickening K-means or Voronoi partitions in Hilbert spaces (see Sect. 3.3). Indeed, in this case, it is possible to test whether a given shortest path intersects a cover element or intersection by computing the intersection of the line induced by the shortest path—which is possible since shortest paths are segments—and all the mediator lines that form the boundary of the cover element.

In practice, when $\hat{\omega}_{\mathrm{SP}}$ and $\ell(\mathbb{X}_n, \hat{f}, \mathcal{U})$ are difficult to compute, we adopt a conservative approach by considering for the graph refinement parameter s the largest possible integer that still allows our estimator to be computed with a reasonable amount of time and memory usage, depending on the machine that is being used. Finally, it should be noted that small sizes of cover elements or intersections induce small ℓ and large s, and thus potentially longer computation times.

3.3 Cover control

In this section, we study the resolutions of covers induced by Voronoi partitions. In particular, we define those covers in Sect. 3.3.1, and provide upper bounds on their resolutions in Sect. 3.3.2. This allows to formulate more explicit upper bounds for the first term in Theorem 3.5.

3.3.1 Defining covers

Covers with hypercubes is the most common cover used with Mappers when $\mathcal{Z} = \mathbb{R}^p$ when p is not too large. When the filter domain $\mathcal{Z}$ is a general length space, as for instance when $\mathcal{Z}$ is the space of probability distributions of $\mathbb{R}$ (see Sect. 4.2) or when $\mathcal{Z}$ is the space of combinatorial graphs (see Sect. 4.3), we need an alternative construction to define covers. A simple way of generating a cover is by using a partition of this space, and thickening the elements of this partition.

Definition 3.6 Let $(\mathcal{Z}, d_{\mathcal{Z}})$ be a length space, and let $\epsilon > 0$.

- For $U \subseteq \mathcal{Z}$ a subset of $\mathcal{Z}$, the *ϵ-thickening* of U is defined as $U^\epsilon = \{z \in \mathcal{Z} : \inf\{d_{\mathcal{Z}}(z, \tilde{z}) : \tilde{z} \in U\} \leq \epsilon\}$.
- Let $\widetilde{\mathcal{Z}}$ be a subset of $\mathcal{Z}$ and let $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ be a partition of $\widetilde{\mathcal{Z}}$, i.e., $\widetilde{\mathcal{Z}} \subseteq \bigcup_{\alpha \in A} U_\alpha$ and $U_\alpha \cap U_\beta = \emptyset$ for all $\alpha \neq \beta \in A$. The *ϵ-thickening* of $\mathcal{U}$ for covering $\widetilde{\mathcal{Z}}$ is defined as $\mathcal{U}^\epsilon = \{U_\alpha^\epsilon\}_{\alpha \in A}$.

Even when $\mathcal{Z} = \mathbb{R}^p$, it might be interesting to use thickenings of partitions instead of hypercube covers, since the number of hypercubes increases exponentially with the dimension p, with many of them having an empty preimage under f, and thus useless.

Note also that when our Mapper-based estimator is used with an ϵ-thickening cover, our estimator gets more difficult to compute when the thickening parameter ϵ goes to zero, since it requires refining the initial neighborhood graph with a lot of new vertices.

3.3.2 Bounding the resolution for ϵ-thickening of Voronoi partitions

Under the same general assumptions as Sect. 3.2, we consider the specific case where $\mathcal{Z} = \mathbb{R}^p$ endowed with the inner product $\langle \cdot, \cdot \rangle$. Partitions of $\mathbb{R}^p$ can be computed very efficiently, for instance with Voronoi partitions and the k-means algorithm. In this section we give an upper bound on the resolution term involved in the upper bound of Theorem 3.5, for a cover computed from a k-means algorithm.

Definition 3.7 For Q a measure in $\mathbb{R}^p$ and $k \in \mathbb{N}^*$, a set $t(Q)$ of k points in $\mathbb{R}^p$ is said to be k-optimal for Q if

$$t(Q) \in \operatorname*{argmin}_{t \in (\mathbb{R}^p)^k} \int_{\mathcal{Z}} \min_{i=1,\dots,k} \|z - t_i\|_2^2 \, dQ(z).$$

Let $P_n^{\hat{f}}$ be the push forward measure of P_n by the stochastic filter function $\hat{f}$, where P_n is the empirical measure associated to $\mathbb{X}_n$. Note that all these quantities are defined conditionally to the sample $\mathbb{X}_n$ and $\hat{f}$ (in particular if $\hat{f}$ depends on other observations $\mathbb{X}'_n$, as would be the case, e.g., in the context of a regression function filter). Note also that $P_n^{\hat{f}}$ is equivalently defined as the empirical measure corresponding to the observation of the sample $\mathbb{Z}_n = \hat{f}(\mathbb{X}_n)$. The k-means algorithm on $\mathbb{Z}_n$ aims at approximating an optimal k points for the empirical measure $P_n^{\hat{f}}$ from the observation $\mathbb{Z}_n$.

Let $\hat{t} = t(P_n^{\hat{f}})$ be an optimal k-points for the empirical measure $P_n^{\hat{f}}$. We denote by $\widehat{\mathcal{U}}^\epsilon = \{\hat{U}_j^\epsilon\}_{j=1,\dots,k}$ the ϵ-thickening of the Voronoi partition associated to $\hat{t}$. Since $\mathcal{Z} = \mathbb{R}^p$, we know that $\hat{f}_{\mathrm{SP}}$ is a linear interpolation between the Z_i's. Thus $\widehat{\mathcal{U}}^\epsilon$ is a cover for $\mathrm{im}(\hat{f}_{\mathrm{SP}})$ and Assumption (H4) is satisfied.

We give our result under the additional assumption that the our modulus of continuity ω is upper bounded by a concave function of the form $\bar{\omega}(u) = cu^\gamma$, with $c \geq 0$ and $0 < \gamma \leq 1$. This assumption is obviously stronger than Assumption (H3).

- **(H5) Power function upper bounds** ω. There exists $\gamma \in (0, 1]$ and $c \in \mathbb{R}^+$ such that for any $u \in \mathbb{R}^+$, $\omega(u) \leq \bar{\omega}(u) = cu^\gamma$.

This technical assumption allows us to provide a simple upper bound. Moreover, upper bounding ω by a concave function makes sense since the *minimal modulus of continuity* ω_f defined with

$$\omega_f(u) = \sup \left\{ d_{\mathcal{Z}}(f(x), f(x')) \, : \, x, x' \in \mathcal{X} \text{ and } \|x - x'\| \leq u \right\}$$

is also a concave function on the compact set $\mathcal{X}$ (see for instance Section 6 in DeVore and Lorentz (1993)), and satisfies $\omega_f \leq \omega$. The next result gives a control on the resolution of $\widehat{\mathcal{U}}^\epsilon$ in $\mathbb{R}^p$ with respect to the filter function $\hat{f}_{\mathrm{SP}}$.

Theorem 3.8 *Under assumptions* (H1), (H2) *and* (H5), *for* $\mathcal{Z} = \mathbb{R}^p$ *and for* $k \leq \frac{n}{p+2}$, *the resolution of the cover* $\widehat{\mathcal{U}}^\epsilon$ *in* $\mathbb{R}^p$ *with respect to the filter function* $\hat{f}_{\mathrm{SP}}$ *satisfies*

$$\mathbb{E}\left[\mathrm{res}(\widehat{\mathcal{U}}^\epsilon, \hat{f}_{\mathrm{SP}})\right] \leq C_1 \left[k^{-\frac{2\gamma^2}{b^2+2\gamma b}} + \left(\frac{kp}{n}\right)^{\frac{\gamma}{2b+4\gamma}} + \mathbb{E}\|(f-\hat{f})|_{\mathbb{X}_n}\|_\infty \right] + 2\varepsilon. \quad (7)$$

Consequently, the following risk bound holds for our Mapper based estimator $\mathrm{M}_n = \mathrm{M}_{\hat{f}_{\mathrm{SP}}, \widehat{\mathcal{U}}^\epsilon, G_{\delta,s}}(\mathbb{X}_{n,s})$*:*

$$\begin{aligned} &\mathbb{E}\left[d_{\mathrm{GH}}((\mathrm{M}_n, \tilde{d}_{\hat{f}_{\mathrm{SP}}, \widehat{\mathcal{U}}^\epsilon}), (\mathrm{R}_f(\mathcal{X}), \tilde{d}_f))\right] \\ &\leq C_2 \left[k^{-\frac{2\gamma^2}{b^2+2\gamma b}} + \left(\frac{kp}{n}\right)^{\frac{\gamma}{2b+4\gamma}} + \mathbb{E}\|(f-\hat{f})|_{\mathbb{X}_n}\|_\infty \right] + 10\varepsilon. \end{aligned} \quad (8)$$

Moreover, the constants C_1 *and* C_2 *depends on a, b, c,* γ, $\|f\|_\infty$ *and on the geometric parameters of* $\mathcal{X}$.

The proof of Theorem 3.8 is given in Sect. B.2, in which several ideas from Brécheteau and Levrard (2020) are reused and adapted. Note that we could also provide a deviation bound on the resolution by applying the so-called Bounded Inequality in a standard way (see for instance Theorem 6.2 in Boucheron et al. (2005)).

Rate of convergence Assuming that γ and b are known, we can choose k to balance the first two terms in the bracket of the right hand side of Inequality (8). By taking k of the order of $\left(\frac{n}{p}\right)^{\frac{b^2+2\gamma b}{(b+2\gamma)(4\gamma+b)}}$, we obtain that the first two terms in the bracket are of the order of $\varepsilon_n := \left(\frac{p}{n}\right)^\zeta$ with $\zeta := \frac{2\gamma^2}{(b+2\gamma)(4\gamma+b)} < \frac{1}{4}$. If the convergence of $\hat{f}$ to f is faster than ε_n, and taking a resolution ε of the order of ε_n, we finally obtain that the expected risk of our Mapper based estimator is of the order of ε_n. We conjecture that this rate of convergence is not optimal, however it can be used to show the consistency of our Mapper-based estimator.

3.4 Application to KPCA filters

In this section, we study the upper bounds of Theorem 3.5 in the particular case where $\mathcal{Z}$ is a reproducing kernel Hilbert space (RKHS). Let $\mathcal{Z}$ be a RKHS associated to a continuous kernel function K defined on $\mathcal{X} \times \mathcal{X}$. The set $\mathcal{X}$ being compact and K being continuous, $\mathcal{Z}$ is then a separable RKHS. Moreover, the feature map $x \mapsto K(x, \cdot)$ is a continuous function from $\mathcal{X}$ to $\mathcal{Z}$ since:

$$\|K(x,\cdot) - K(x',\cdot)\|_{\mathcal{Z}}^2 = K(x,x) + K(x',x') - K(x,x') - K(x',x). \quad (9)$$

Moreover, the random variable $Z = K(X, \cdot)$ is bounded in $\mathcal{Z}$ since $\|K(X,\cdot)\|_{\mathcal{Z}}^2 = K(X,X)$ which is almost surely bounded on the compact space $\mathcal{X}$. In particular, $\mathbb{E}\left[\|K(X,\cdot)\|_{\mathcal{Z}}^2\right] < \infty$. In this setting, the covariance operator of the distribution of Z is well defined (see for instance Section 2 and 4.1 in Blanchard et al. (2007)). To simplify, we will assume that the distribution of Z is centered.

Covariance operator Let $\Gamma = \mathbb{E}(Z \otimes Z^*)$ be the covariance operator and let Π_p be the orthogonal projection operator on the set of the first p eigenvectors of Γ. The operator Γ can be approximated by its empirical version:

$$\Gamma_n = \frac{1}{n}\sum_{i=1}^{n} Z_i \otimes Z_i^*$$

where $Z_i = K(X_i, \cdot)$. Let $\hat{\Pi}_{n,p}$ be the orthogonal projection operator on the set of the first p eigenvectors of Γ_n. In this section, we consider as filter functions the composition of the feature map with one of the two projection operators:

$$f_p : x \in \mathcal{X} \mapsto \Pi_p\left(K(x,\cdot)\right)$$

and

$$\hat{f}_{n,p} : x \in \mathcal{X} \mapsto \hat{\Pi}_{n,p}\left(K(x,\cdot)\right).$$

Modulus of continuity Let ω_K be the modulus of continuity of K: for any $x_1, x_2, x_1', x_2' \in \mathcal{X}$,

$$\left|K(x_1,x_2) - K(x_1',x_2')\right| \leq \omega_K\left(\sqrt{\|x_1 - x_1'\|^2 + \|x_2 - x_2'\|^2}\right),$$

where $\|\cdot\|$ is the euclidean norm of $\mathbb{R}^D$. Let $x, x' \in \mathcal{X}$, then

$$\begin{aligned}\left\|f_p(x) - f_p(x')\right\|_{\mathcal{Z}} &= \left\|\Pi_p\left(K(x,\cdot)\right) - \Pi_p\left(K(x',\cdot)\right)\right\|_{\mathcal{Z}} \\ &\leq \left\|K(x,\cdot) - K(x',\cdot)\right\|_{\mathcal{Z}} \\ &\leq \sqrt{2\omega_K(\|x - x'\|)}\end{aligned}$$

where the last inequality comes from (9). This shows that $\sqrt{2\omega_K}$ is a modulus of continuity for f_p.

Upper bound The statistical analysis of PCA in Hilbert spaces has been the subject of several works, see for instance (Reiß and Wahl 2020; Blanchard et al. 2007; Biau and Mas 2012; Shawe-Taylor et al. 2005). Here, we need a control for the sup norm between the filter and its empirical version. According to Theorem 2.1 in Biau and Mas (2012):

$$\mathbb{E}\left[\sup_{z\in\mathcal{Z},\,\|z\|_{\mathcal{Z}}\leq 1} \|\Pi_p(z) - \hat{\Pi}_{n,p}(z)\|_{\mathcal{Z}}\right] \leq \frac{C}{\sqrt{n}}$$

where the constant C only depends on p. Since $\mathcal{X}$ is compact and $x \mapsto K(x, \cdot)$ is continuous, it follows that:

$$\mathbb{E}\left[\sup_{x \in \mathcal{X}} \| f_p(x) - \hat{f}_{n,p}(x) \|_{\mathcal{Z}}\right] \leq \frac{C'}{\sqrt{n}}$$

where C' depends on $D_{\mathcal{X}}$ and p. Under assumptions (H1), (H2) and (H5), if we perform a k-means algorithm in the space of the p first components of the KPCA to derive a cover as explained in Sect. 3.3.2, we can then apply Theorem 3.8 to our corresponding Mapper-based estimator. The convergence of the estimated filter in $O(1/\sqrt{n})$ is fast enough so that it does not slow down the convergence of our Mapper-based estimator. For k and ε chosen as in the discussion following Theorem 3.8, we finally obtain that the risk of our Mapper based estimator can be upper bounded by a term of the order of $\left(\frac{p}{n}\right)^{\frac{2\gamma^2}{(b+2\gamma)(4\gamma+b)}}$.

4 Applications of Mapper in the Stochastic Filter setting

In this section, we focus on examples and applications of the Stochastic Filter setting (see Sect. 3), in which the filter $\hat{f}$ used to compute the Mapper is assumed to be an estimation (computed from the data sample) of the true target filter f used to compute the Reeb space. We first provide in Sect. 4.1 various examples of stochastic filters in statistics and machine learning. Indeed, standard methods provide estimated regression functions and classification probability estimates which are interesting to study with Mapper. Then, we turn the focus to the length space of probability distributions in Sect. 4.2, and we finally provide an illustration for the length space of combinatorial graphs with the graph edit distance in Sect. 4.3. Throughout this section, the Mappers that are computed and discussed always refer to our Mapper-based estimator.

4.1 Stochastic Filter in Statistical Machine Learning

In this section, we discuss the various potential applications of Mappers in statistical machine learning, in which the filter is often used for inference and prediction, and we provide associated numerical experiments and illustrations. We also refer the interested reader to Hastie et al. (2003) for more details on the statistical and machine learning methods used in this section.

Stochastic real-valued filters We first consider a few applications in which the estimated and true target filters are real-valued functions, i.e., $\mathcal{Z} = \mathbb{R}$. In this setting, one can apply either the risk bound given in Theorem 3.5 or the results from Carrière et al. (2018) to quantify the approximation and convergence of Mapper.

- **Inference** When the target filter function only depends on the measure P itself, we can define estimators of this filter using the point cloud $\mathbb{X}_n$ alone. For instance, a dimension reduction filter (e.g. PCA), the eccentricity filter or the density estimator

filter are all estimators of underlying filters defined from P. See for instance (Carrière et al. 2018) for examples.

- **Regression** We now assume that we observe a random variable Y_i at each point X_i:

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n \tag{10}$$

where the true filter is $f(x) = \mathbb{E}(Y|X = x)$, i.e., the regression function on $\mathcal{X}$ and $\varepsilon_i = Y_i - f(X_i)$. Then, the Mapper of $\mathbb{X}_n$ can be computed with any estimator $\hat{f}$ of f (from the statistical regression literature) in order to infer the Reeb space $\mathrm{R}_f(\mathcal{X})$.

- **Binary classification** We now assume that we observe a binary variable $Y_i \in \{-1, 1\}$ at each point X_i of the sample. Let $f(x) = P(Y = 1|X = x)$ be the probability of class 1 for any $x \in \mathcal{X}$. In this setting, inferring the target Reeb space $\mathrm{R}_f(\mathcal{X})$ with a Mapper computed on $\mathbb{X}_n$ for some estimator $\hat{f}$ of the class probability distribution (given by any machine learning classifier) would provide insights about how data is topologically stratified w.r.t. the confidence given by the classifier.

Extension to stochastic multivariate filters For many problems in statistical machine learning, the quantity of interest is actually a multivariate quantity. In this setting, using Theorem 3.5 allows to statistically control the quality of Mapper, which, to our knowledge, is new in the Mapper literature.

- **Dimension reduction** In this setting, a natural extension of real-valued inference described above is the projection onto the p first directions of any dimension reduction algorithm. The corresponding Mapper is now a multivariate Mapper and the underlying filter is the projection onto the p first directions of the covariance operator of P. See Sect. 3.4 above.
- **Multivariate regression** Multivariate regression is the generalization of (univariate) regression when the variable Y in Eq. (10) is now a random vector.
- **Multi-class classification** We observe a categorical variable $Y_i \in \{0, \ldots, k\}$ at each point X_i. Let $f_k(x) = P(Y = k|X = x)$ be the probability of class k at $x \in \mathcal{X}$. The underlying filter is now the vector of estimated probabilities $f = (f_0, \ldots, f_k)$, which can be obtained with classification methods in statistical machine learning.

Synthetic example We now describe two multi-class classification problems and display the corresponding Mappers. In the first one, we generated a data set in two dimensions with three different classes which are entangled with each other. See Fig. 2 (left) for an illustration. We then trained a Random Forest classifier on this data set, and computed the estimated probabilities for each of the training points, meaning that we have an estimated multivariate filter $\hat{f} : \mathbb{R}^2 \to [0, 1]^3$. The corresponding Mapper (computed with 10 intervals and overlap 30% for each class) is shown in Fig. 2 (right). Moreover, the Mapper nodes are colored with the variance of the class probability distributions: the smaller the variance, the more confident the prediction. It is clear from the Mapper that the classifier induces a topological stratification of the data, in the sense that points in the middle of the space (located in the middle of the triangle-shaped Mapper), on which the classifier is unsure, connect with points for

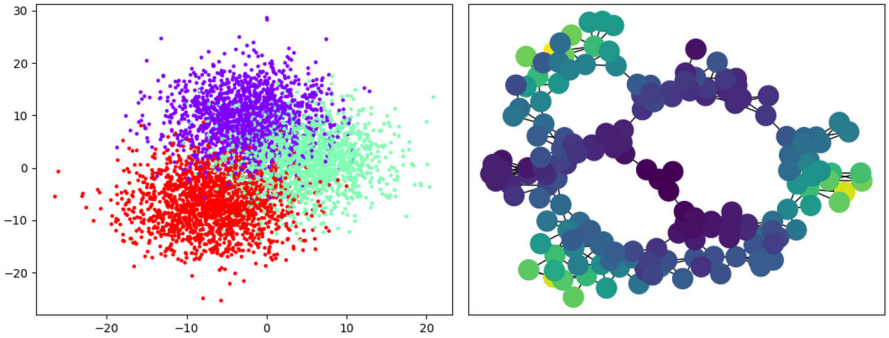

Fig. 2 Three label classification problem and its corresponding Mapper. Left: we generate points in 2D with three different groups (red, purple, green). Right: Mapper computed with the posterior probability of a Random Forest classifier. Nodes are colored with the variance of the estimated probabilities from low (yellow) to large (dark blue)

which the classifier hesitates between two classes (located in the middle of the "edges" of the triangle), which themselves connect with points where the classifier is confident (located at the "corners" of the triangle), leading to some non-trivial 1-dimensional topological features (i.e., circles) in the data[3], which are not visible at first sight on the data set. We believe this visualization could be of great help when it comes to interpreting the output of standard statistical machine learning methods.

Accelerometer data In our second example, we study a data set of time series obtained from accelerometers placed on people doing six possible types of activities, namely "standing", "sitting", "laying", "walking", "walking upstairs" and "walking downstairs". From the raw data, 561 features have been extracted from sliding window, see (Anguita et al. 2013) and the data website[4] for more details. A Naive Bayes classifier has been trained on the 7, 352 observations. We finally generated an associated Mapper with the corresponding estimated probabilities (computed with 3 intervals and 30% gain for each class), and we colored the nodes with variance, similarly to what was done above. We show the Mapper, as well as representative time series for some of its nodes, in Fig. 3. Again, the classifier is inducing a topological stratification of the data, with two connected components (corresponding to the two global types of activities, namely walking activities or stationary activities), which are themselves stratified into three activities connected by time series where the classifier is unsure.

4.2 Stochastic Filter with Conditional Probability Distributions

We now assume that we observe an i.i.d sample $\{(X_i, Y_i) : 1 \leq i \leq n\}$, where $X_i \in \mathcal{X}$ and $Y_i \in \mathbb{R}$. In this setting, we propose to consider the more complex filter function which is defined as the conditional distribution $(Y|X)$: the value of this filter at x is

[3] One might ask whether these circles are artifacts from graph visualization. We checked that these circles are intrinsic to the structure by confirming their presence in the 1-dimensional homology group of the complex, that we computed manually.

[4] https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones.

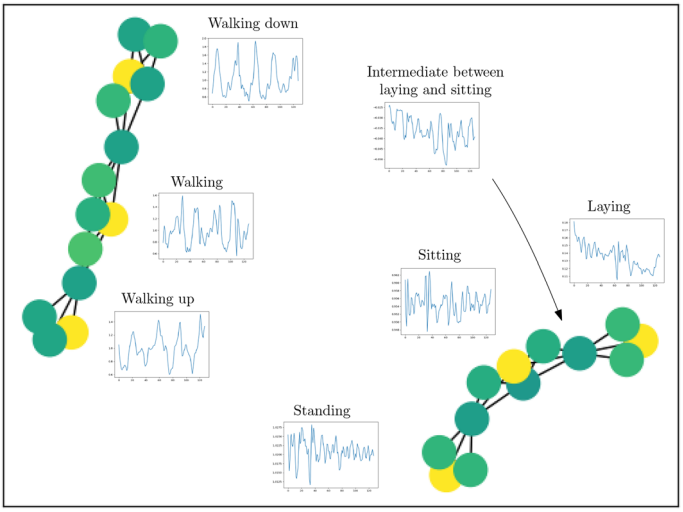

Fig. 3 Mapper computed on accelerometer data with the posterior probability of a Naive Bayes classifier. Nodes are colored with the variance of the estimated probabilities from low (yellow) to large (dark green)

the conditional distribution $(Y|X)$. In this framework, the filter domain $\mathcal{Z}$ is thus the space of probability distributions.

In practice, it might be tempting to directly compute the standard Mapper with the Y_i's as filter values. However, this approach does not really make sense because there is no reason for this Mapper to converge to a deterministic Reeb space for some underlying filter function. Having in mind that the relevant target filter is the conditional probability distribution $(Y|X)$, it is clear that this naive approach is not a good strategy for this aim, since single observations can be very poor estimates of the corresponding distributions.

4.2.1 Mapper with probability distributions

Let $\mathcal{P}$ be the set of probability measures on $\mathbb{R}$. For $x \in \mathcal{X}$, let $\succeq_x$ be the conditional distribution $(Y|X = x)$. Let ν be the filter $\nu : x \in \mathcal{X} \mapsto \nu_x \in \mathcal{P}$. Various metrics can be proposed on $\mathcal{P}$, one of them being the Prokhorov metric (Billingsley 2013), which metrizes weak convergence. Generally speaking, the Reeb space $\mathrm{R}_\nu(\mathcal{X})$ is difficult to infer since it requires to estimate the conditional probability distribution $\succeq_x$ for all points of $\mathcal{X}$, which is a difficult task, especially for high dimensional data—see for instance (Efromovich 2007). As far as we know, conditional density estimation on submanifolds has not been studied yet. Moreover, as soon as ν is injective, which is not a strong assumption in practice, the Reeb space will be isomorphic to $\mathcal{X}$ and it will not provide more information than standard manifold learning procedures (Ma and Fu 2011). We thus propose to study approximations of $\mathrm{R}_\nu(\mathcal{X})$, using a filter that is a simple descriptor (such as the mean or the histogram) of ν_x. In this situation, from a data analysis perspective, crude approximations of the Reeb space shows more interesting patterns than those provided by the Reeb space itself.

Mean- and histogram-based Mappers Let $\mathcal{I} = (I_1, \ldots, I_d)$ be a partition of $\mathbb{R}$ with intervals. We define the histogram filter Hist associated to $\mathcal{I}$ by $\text{Hist}_j(x) = P\left(Y \in I_j \,|X = x\right)$

for $j = 1, \ldots, d$. The codomain of Hist is in $\mathbb{R}^d$, i.e., it is a multivariate filter, with corresponding Reeb space $\text{R}_{\text{Hist}}(\mathcal{X})$. We then propose to compute the Mapper with an estimated histogram, which we call the *histogram-based Mapper*, using the Nadaraya-Watson kernel estimator:

$$\widehat{\text{Hist}}_j(x) = \frac{\sum_{i=1,\ldots,n} \mathbb{1}_{Y_i \in I_j} K_h(X_i - x)}{\sum_{i=1,\ldots,n} K_h(X_i - x)}$$

where $K_h(x) = \frac{1}{h} K(\frac{x}{h})$ for a kernel function K, which we choose, in practice, to be the indicator function of the unit ball in the ambient Euclidean space.

Note that a simpler approach is to estimate the (conditional) mean $f(x) = \mathbb{E}(Y|X = x)$, and we call the corresponding estimator the *mean-based Mapper*. However, as illustrated in numerical experiments presented below, it may be not sufficient to retrieve interesting data structure.

4.2.2 Numerical experiments

We now provide examples of computations of our Mapper-based estimators computed from single realizations of synthetic conditional probability distributions [5]. We generate 5, 000 points from an annulus, and we looked at two conditional distributions for each point, namely Gaussians and bimodal ones. See Figs. 4 and 5. In each of these figures, we display five Mappers: the standard Mapper, the mean-based Mapper when the true conditional mean is supposed to be known, the mean-based Mapper when this mean is estimated, the histogram-based Mapper when the true histogram is supposed to be known, and the histogram-based Mapper when the histogram is estimated. We also plot, for the standard Mapper and the mean-based Mappers, a 3D embedding of the data set, with the mean values used as height. For the standard Mapper and the mean-based Mappers, we used an interval cover with 15 intervals and overlap percentage 30%. For the histogram-based Mapper, we used histograms with 100 bins and an 0.5-thickening of a K-PDTM cover (Brécheteau and Levrard 2020) with $K = 10$ cover elements.

Gaussian conditional In Fig. 4, we generated Gaussian conditional probability distributions centered on the second coordinates of the points. It can be seen that the standard Mapper recovers the underlying structure, but in a very imprecise way, in the sense that the feature size is much smaller than it should be, due to the variances of the distributions that induce very noisy filter values. On the other hand, the mean-based Mappers and the histogram-based Mappers all recover the correct structure in much more precise fashion.

Bimodal conditional In Fig. 5, we generate bimodal conditional probability distributions whose modes are centered on the second coordinate and its opposite (minus the

[5] Our code is freely available at https://github.com/MathieuCarriere/metricmapper.

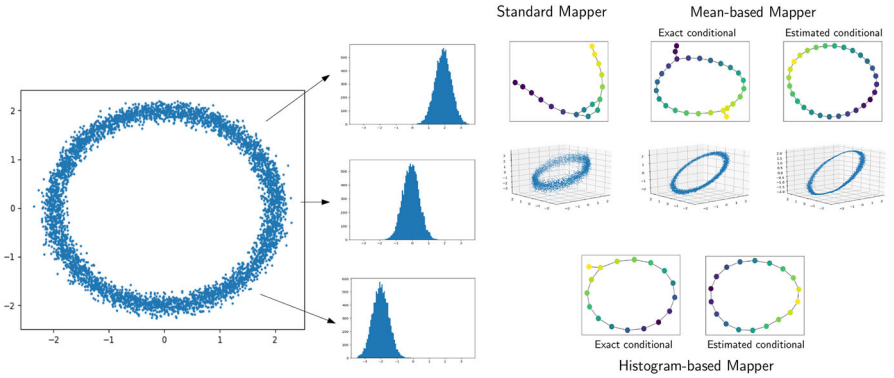

Fig. 4 Standard, mean- and histogram-based Mappers computed with exact and estimated Gaussian conditional probability distributions

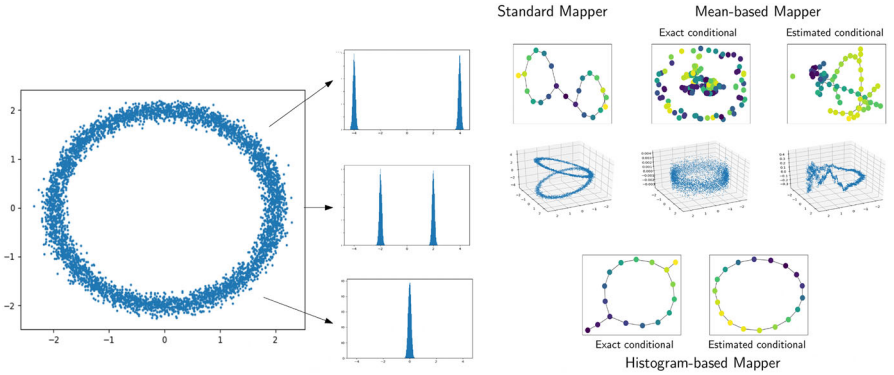

Fig. 5 Standard, mean- and histogram-based Mappers computed with exact and estimated bimodal conditional probability distributions

minimum of the coordinates values). This way, all conditional probability distributions have the same mean. This time, the standard Mapper gets fooled by the probability distributions, and outputs two topological structures instead of one, due to the two modes of the distributions. The mean-based Mappers also fail due to the fact that the distributions all have the same mean, which mixes all points together and makes topological inference very difficult, leading to very noisy Mappers. On the other hand, the histogram-based Mappers both manage to retrieve the correct structure in a precise way.

4.3 Stochastic filter with combinatorial graphs

We end this application section by providing an example of our Mapper-based estimator, when the domain of the filter function is the space of combinatorial graphs. More specifically, we generated a graph for each data point of the annulus data set, using the Erdős-Rényi model on 20 nodes, and using the first coordinate of the points

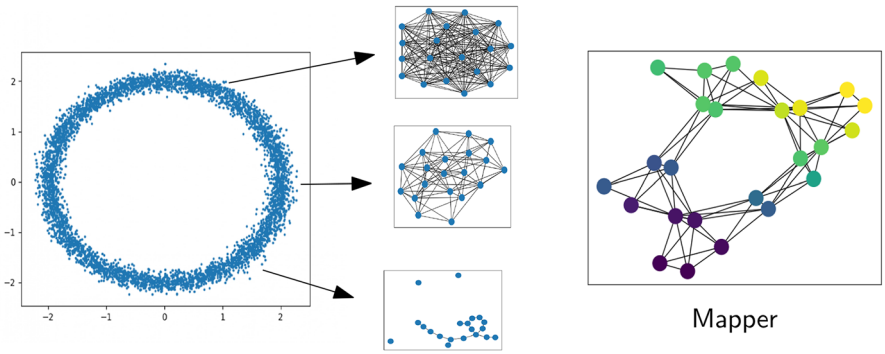

Fig. 6 Example of Mapper computation for combinatorial graphs

(normalized between 0 and 1) as the model parameter (that is, any possible edge among the 20 nodes appears with probability given by the model parameter). This means that points located at the bottom of the annulus will have graphs with fewer edges than those above. See Fig. 6 (left). Then, we used the graph edit distance (provided in the `networkx` Python package) and a Voronoi cover with 10 cells (corresponding to 10 randomly sampled germs) and 0.5-thickening to compute our estimator. The corresponding Mapper is shown in Fig. 6 (right). One can see that the correct topology is retrieved by our estimator.

5 Conclusion and future directions

In this article, we presented a computable Mapper-based estimator that enjoys statistical guarantees for its approximation of its corresponding target Reeb space. Moreover, we demonstrated how it can be applied when the filter is estimated from a random sample of data, which we call the Stochastic Filter setting. In this case, we demonstrated a few applications in statistical machine learning, and we provided examples in which the usual Mapper fails dramatically, whereas our estimators still succeed. Much work is still needed for future directions, including demonstrating optimality and stability of the estimator. Moreover, we plan on adapting bootstrap methods to compute and interpret confidence regions. We also plan to adapt specific clustering algorithms in the space of distributions to propose efficient covers in this setting. In the longer term, we also plan to strengthen our results by extending them to the interleaving distance of Munch and Wang (2016).

Acknowledgements The authors would like to thank Claire Brécheteau and Clément Levrard for helpful discussions on the control of the resolution of the k-means algorithm, and Yusu Wang for suggesting the use of filter-based pseudometrics.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

A Reach of a manifold

Let $\mathcal{X}$ be a compact submanifold of $\mathbb{R}^D$. Its *medial axis* is the subset $\Gamma(\mathcal{X})$ defined with $\Gamma(\mathcal{X}) := \{x \in \mathbb{R}^D \,:\, \mathrm{NN}_{\mathcal{X}}(x) \geq 2\}$, where $\mathrm{NN}_{\mathcal{X}}(x)$ denotes the set of closest points of x in $\mathcal{X}$. The *reach* of $\mathcal{X}$ is the distance of $\mathcal{X}$ to its medial axis, i.e., the quantity $\mathrm{rch}(\mathcal{X}) := \inf\{\|x-g\| \,:\, x \in \mathcal{X}, g \in \Gamma(\mathcal{X})\}$. Intuitively, the reach is related to curvature, and quantifies how smooth a manifold is.

B Proofs

B.1 Proof of Theorem 3.5

We assume that (H1), (H2), (H3) and (H4) of Sect. 3.2 are satisfied. The parameters $s \geq s_n$ and δ_n are assumed to be chosen according to (5) and (6). Recall that the point cloud $\mathbb{X}_{n,s}$ is a refinement of the point cloud $\mathbb{X}_n$, as defined in Sect. 3.1. We also introduce the generalized inverse of a modulus of continuity ω:

$$\omega^{-1}(v) = \{u \,:\, \omega(u) \geq v\}\,.$$

Approximation Lemmata We first prove three approximations. In the first one, we show that our estimator M_n is actually equivalent to the (continuous) Mapper of an associated neighborhood graph.

Lemma B.1 *The Mappers* M_n *and* $\mathrm{M}_{\hat{f}_{\mathrm{SP}},\mathcal{U}}(G_{\delta_n})$ *are isomorphic as simplicial complexes. Hence,*

$$d_{\mathrm{GH}}\left((\mathrm{M}_n, \tilde{d}_{\hat{f}_{\mathrm{SP}},\mathcal{U}}), (\mathrm{M}_{\hat{f}_{\mathrm{SP}},\mathcal{U}}(G_{\delta_n}), \tilde{d}_{\hat{f}_{\mathrm{SP}},\mathcal{U}})\right) = 0$$

Proof Let $U_\alpha \in \mathcal{U}$, and $\mathcal{C}_\alpha$ be a connected component of $\hat{f}_{\mathrm{SP}}^{-1}(U_\alpha)$ in G_{δ_n}. We claim that $\mathcal{C}_\alpha \cap \mathbb{X}_{n,s} \neq \emptyset$. Indeed, if we assume that $\mathcal{C}_\alpha \cap \mathbb{X}_{n,s} = \emptyset$ then it means that $\mathcal{C}_\alpha$ is constituted from a subpath $\bar{e}$ of an edge e of G_{δ_n}, that does not contain the endpoints of e in $\mathbb{X}_n$, nor any points of $\mathbb{X}_{n,s}$ in the subdivision of e. Hence, e is element-crossing and U_α is crossed by e. By definition, the length $|\hat{f}_{\mathrm{SP}}(e) \cap U_\alpha|$ must be at least $\ell(\mathbb{X}_n, \hat{f}, \mathcal{U})$. Moreover, due to the subdivision process, the length $|\bar{e}|$ must be less than $\delta_n/(1+s) \leq \delta_n/(1+s_n)$, meaning that $|\hat{f}_{\mathrm{SP}}(\bar{e})| = |\hat{f}_{\mathrm{SP}}(e) \cap U_\alpha|$ must be less than $\hat{\omega}_{\mathrm{SP}}(\delta_n/(1+s_n))$. Hence, using the definition of s_n, we have the following inequalities:

$$\frac{\ell}{2} \geq \hat{\omega}_{\mathrm{SP}}\left(\frac{\delta_n}{1+\left\lfloor \delta_n/\hat{\omega}_{\mathrm{SP}}^{-1}(\ell/2)\right\rfloor}\right) \geq |\hat{f}_{\mathrm{SP}}(e) \cap U_\alpha| \geq \ell,$$

which leads to a contradiction (except for $\ell = 0$, which happens with null probability).

Hence, for each U_α and connected component $\mathcal{C}_\alpha$ of $\hat{f}_{\mathrm{SP}}^{-1}(U_\alpha)$ in G_{δ_n}, there is one point of $\mathbb{X}_{n,s_n}$ that belongs to $\mathcal{C}_\alpha$. Now, let $\tilde{\mathcal{C}}_\alpha$ be the connected component in $G_{\delta_n,s_n}(U_\alpha)$

(see Equation (1)) associated to this point. We now claim that $\tilde{\mathcal{C}_\alpha}$ is included in $\mathcal{C}_\alpha$. Indeed, since G_{δ_n,s_n} is nothing but a subdivision of G_{δ_n}, and since any edge of G_{δ_n,s_n} in $\tilde{\mathcal{C}_\alpha}$ must also be present in $\mathcal{C}_\alpha$ (otherwise it would induce an element-crossing edge in G_{δ_n} whose intersection with the corresponding crossed cover element would contain no points in $\mathbb{X}_{n,s_n}$, which is impossible for the reason mentioned above), it follows that $\mathcal{C}_\alpha$ deform-retracts on $\tilde{\mathcal{C}_\alpha}$. Hence, M_n and $\mathrm{M}_{\hat{f}_{\mathrm{SP}},\mathcal{U}}(G_{\delta_n})$ have the exact same sets of nodes.

The same argument applies straightforwardly to show that the connected components in the intersections are also in bijection, which means that the simplices of both Mappers are in correspondence as well. □

Let d_g denote the geodesic distance on $\mathcal{X}$. Let $d_{\mathrm{H},n} = d_{\mathrm{H}}^g(\mathbb{X}_n, \mathcal{X}) = \sup_{x\in\mathcal{X}} \min_{X\in\mathbb{X}_n} d_g(x, X)$ denote the Hausdorff distance between $\mathcal{X}$ and $\mathbb{X}_n$ computed with geodesic distances.

Lemma B.2 *Let $x, x' \in G_\delta$. Then, $|d_{f_{\mathrm{SP}}}(x, x') - d_{f_{\mathrm{SP}}}(\zeta(x), \zeta(x'))| \leq 2\omega(\delta)$, where $\zeta(x) \in \mathbb{X}_n$ is the closest endpoint of the edge to which x belongs if $x \notin \mathbb{X}_n$ and x otherwise.*

Similarly, let $x, x' \in \mathcal{X}$. Then, $|d_f(x, x') - d_f(\zeta(x), \zeta(x'))| \leq 2\omega(d_{\mathrm{H},n})$, where $\zeta(x) \in \mathbb{X}_n$ is such that $d_g(x, \zeta(x)) \leq d_{\mathrm{H},n}$ (whose existence is guaranteed with $d_{\mathrm{H}}^g(\mathbb{X}_n, \mathcal{X}) = d_{\mathrm{H},n}$).

Proof Let γ be a path going from $\zeta(x)$ to $\zeta(x')$ achieving[6] $d_{f_{\mathrm{SP}}}(\zeta(x), \zeta(x'))$. Let $\gamma' = \gamma_{x'} \circ \gamma \circ \gamma_x$, where γ_x is the path going from x to $\zeta(x)$ along the edge e_x to which x belongs, and $\gamma_{x'}$ is the path going from $\zeta(x')$ to x' along the edge $e_{x'}$ to which x' belongs. Also, let $\tilde{\zeta}(x) \in \mathbb{X}_n$ denote the other endpoint of e_x, and similarly for $\tilde{\zeta}(x')$. Then γ' is a path from x to x'.

Now, $f_{\mathrm{SP}} \circ \gamma' = f_{\mathrm{SP}} \circ \gamma_{x'} \cup f_{\mathrm{SP}} \circ \gamma \cup f_{\mathrm{SP}} \circ \gamma_x \subseteq f_{\mathrm{SP}} \circ e_{x'} \cup f_{\mathrm{SP}} \circ \gamma \cup f_{\mathrm{SP}} \circ e_x$, and

$$
\begin{aligned}
&\mathrm{diam}_{\mathcal{Z}}(f_{\mathrm{SP}} \circ \gamma') \leq \mathrm{diam}_{\mathcal{Z}}(f_{\mathrm{SP}} \circ e_{x'} \cup f_{\mathrm{SP}} \circ \gamma \cup f_{\mathrm{SP}} \circ e_x) \\
&\leq \max\{d_{\mathcal{Z}}(f(u), f(v)) : u, v \in \{\tilde{\zeta}(x), \tilde{\zeta}(x')\} \cup (\gamma \cap \mathbb{X}_n)\} \\
&\leq \max\{d_{\mathcal{Z}}(f(u), f(v)) : u, v \in (\gamma \cap \mathbb{X}_n)\} + d_{\mathcal{Z}}(f \circ \zeta(x), f \circ \tilde{\zeta}(x)) \\
&\quad + d_{\mathcal{Z}}(f \circ \zeta(x'), f \circ \tilde{\zeta}(x')) \\
&= d_{f_{\mathrm{SP}}}(\zeta(x), \zeta(x')) + d_{\mathcal{Z}}(f \circ \zeta(x), f \circ \tilde{\zeta}(x)) + d_{\mathcal{Z}}(f \circ \zeta(x'), f \circ \tilde{\zeta}(x')) \\
&\leq d_{f_{\mathrm{SP}}}(\zeta(x), \zeta(x')) + 2\omega(\delta)
\end{aligned}
$$

Hence $d_{f_{\mathrm{SP}}}(x, x') \leq d_{f_{\mathrm{SP}}}(\zeta(x), \zeta(x')) + 2\omega(\delta)$.

Now, assume $d_{f_{\mathrm{SP}}}(x, x') < d_{f_{\mathrm{SP}}}(\zeta(x), \zeta(x')) - 2\omega(\delta)$, and let γ be a path from x to x' achieving $d_{f_{\mathrm{SP}}}(x, x')$. Again, let $\gamma' = \gamma_{x'} \circ \gamma \circ \gamma_x$, where γ_x is the path going from $\zeta(x)$ to x along the edge e_x to which x belongs, and $\gamma_{x'}$ is the path going from

[6] We assume for sake of simplicity in this proof that the infimums in the definition of the filter-based pseudometric are always achieved by some path. However, our proof extends straightforwardly to the general case by considering limits of sequences of paths converging to the infimum.

x' to $\zeta(x')$ along the edge $e_{x'}$ to which x' belongs. Then:

$$\begin{aligned}\operatorname{diam}_{\mathcal{Z}}(f_{\mathrm{SP}} \circ \gamma') &\leq \operatorname{diam}_{\mathcal{Z}}(f_{\mathrm{SP}} \circ \gamma_{x'} \cup f_{\mathrm{SP}} \circ \gamma \cup f_{\mathrm{SP}} \circ \gamma_x) \\ &\leq d_{f_{\mathrm{SP}}}(x, x') + d_{\mathcal{Z}}(f(x'), f \circ \zeta(x')) + d_{\mathcal{Z}}(f(x), f \circ \zeta(x)) \\ &\leq d_{f_{\mathrm{SP}}}(x, x') + 2\omega(\delta) \\ &< d_{f_{\mathrm{SP}}}(\zeta(x), \zeta(x')),\end{aligned}$$

which is impossible since $d_{f_{\mathrm{SP}}}(\zeta(x), \zeta(x')) \leq \operatorname{diam}_{\mathcal{Z}}(f_{\mathrm{SP}} \circ \gamma')$.

The proof for the second statement is exactly the same. □

In our third lemma, we show that the Reeb space of a space and its neighborhood graph approximation are actually close, provided that the graph is built on top of a dense enough point cloud.

Lemma B.3 *Assume* $6d_{\mathrm{H},n} \leq \delta_n \leq 2 \cdot \operatorname{rch}(\mathcal{X})$. *Then, one has*

$$d_{\mathrm{GH}}\left((\mathrm{R}_{\hat{f}_{\mathrm{SP}}}(G_{\delta_n}), \tilde{d}_{\hat{f}_{\mathrm{SP}}}), (\mathrm{R}_f(\mathcal{X}), \tilde{d}_f)\right) \leq 4\omega(2\delta_n) + 2\|f_{\mathrm{SP}} - \hat{f}_{\mathrm{SP}}\|_\infty.$$

Proof By the triangle inequality, we have:

$$\begin{aligned}&d_{\mathrm{GH}}\left((\mathrm{R}_{\hat{f}_{\mathrm{SP}}}(G_{\delta_n}), \tilde{d}_{\hat{f}_{\mathrm{SP}}}), (\mathrm{R}_f(\mathcal{X}), \tilde{d}_f)\right) \\ &\leq d_{\mathrm{GH}}\left((\mathrm{R}_{\hat{f}_{\mathrm{SP}}}(G_{\delta_n}), \tilde{d}_{\hat{f}_{\mathrm{SP}}}), (G_{\delta_n}, d_{\hat{f}_{\mathrm{SP}}})\right) \\ &\quad + d_{\mathrm{GH}}\left((G_{\delta_n}, d_{\hat{f}_{\mathrm{SP}}}), (\mathcal{X}, d_f)\right) + d_{\mathrm{GH}}\left((\mathcal{X}, d_f), (\mathrm{R}_f(\mathcal{X}), \tilde{d}_f)\right) \\ &= d_{\mathrm{GH}}\left((G_{\delta_n}, d_{\hat{f}_{\mathrm{SP}}}), (\mathcal{X}, d_f)\right) \text{ by Proposition 2.9} \\ &\leq d_{\mathrm{GH}}\left((G_{\delta_n}, d_{\hat{f}_{\mathrm{SP}}}), (G_{\delta_n}, d_{f_{\mathrm{SP}}})\right) + d_{\mathrm{GH}}\left((G_{\delta_n}, d_{f_{\mathrm{SP}}}), (\mathcal{X}, d_f)\right)\end{aligned}$$

First term Let us bound $d_{\mathrm{GH}}((G_{\delta_n}, d_{\hat{f}_{\mathrm{SP}}}), (G_{\delta_n}, d_{f_{\mathrm{SP}}}))$. The identity map $\iota : G_{\delta_n} \to G_{\delta_n}$ can be used to define a correspondence, with which it is clear that:

$$|d_{f_{\mathrm{SP}}}(x, x') - d_{\hat{f}_{\mathrm{SP}}}(\iota(x), \iota(x'))| = |d_{f_{\mathrm{SP}}}(x, x') - d_{\hat{f}_{\mathrm{SP}}}(x, x')| \leq 2\|f_{\mathrm{SP}} - \hat{f}_{\mathrm{SP}}\|_\infty.$$

Indeed, one has $d_{\hat{f}_{\mathrm{SP}}}(x, x') \leq \operatorname{diam}_{\mathcal{Z}}(\hat{f}_{\mathrm{SP}} \circ \gamma)$, where γ is a path achieving $d_{f_{\mathrm{SP}}}(x, x')$. Thus, $d_{\hat{f}_{\mathrm{SP}}}(x, x') \leq \operatorname{diam}_{\mathcal{Z}}(\hat{f}_{\mathrm{SP}} \circ \gamma) \leq \operatorname{diam}_{\mathcal{Z}}(f_{\mathrm{SP}} \circ \gamma) + 2\|f_{\mathrm{SP}} - \hat{f}_{\mathrm{SP}}\|_\infty = d_{f_{\mathrm{SP}}}(x, x') + 2\|f_{\mathrm{SP}} - \hat{f}_{\mathrm{SP}}\|_\infty$. Symmetrically, one can also show that $d_{f_{\mathrm{SP}}}(x, x') \leq d_{\hat{f}_{\mathrm{SP}}}(x, x') + 2\|f_{\mathrm{SP}} - \hat{f}_{\mathrm{SP}}\|_\infty$, hence the result.

Second term Let us now bound $d_{\mathrm{GH}}((G_{\delta_n}, d_{f_{\mathrm{SP}}}), (\mathcal{X}, d_f))$. Let $\mathcal{C}$ be a correspondence between $\mathcal{X}$ and G_{δ_n} defined with $\mathcal{C} = \{(x, \zeta(x)) : x \in \mathcal{X}\} \cup \{(\zeta(y), y) : y \in G_{\delta_n}\}$ (see Lemma B.2).

Restriction to point cloud First, we show that we can restrict to pairs of points in $\mathbb{X}_n$, up to some constant. Indeed, it follows from Lemma B.2 that, for all $x, x' \in \mathcal{X}$ and $y, y' \in G_{\delta_n}$:

$$\begin{aligned}
|d_f(x,x') - d_{f_{\mathrm{SP}}}(\zeta(x),\zeta(x'))| &\leq |d_f(x,x') - d_f(\zeta(x),\zeta(x'))| + |d_f(\zeta(x),\zeta(x')) \\
&\quad - d_{f_{\mathrm{SP}}}(\zeta(x),\zeta(x'))| \\
&\leq 2\omega(d_{\mathrm{H},n}) + |d_f(\zeta(x),\zeta(x')) - d_{f_{\mathrm{SP}}}(\zeta(x),\zeta(x'))| \\
|d_f(\zeta(y),\zeta(y')) - d_{f_{\mathrm{SP}}}(y,y')| &\leq |d_{f_{\mathrm{SP}}}(y,y') - d_{f_{\mathrm{SP}}}(\zeta(y),\zeta(y'))| + |d_f(\zeta(y),\zeta(y')) \\
&\quad - d_{f_{\mathrm{SP}}}(\zeta(y),\zeta(y'))| \\
&\leq 2\omega(\delta_n) + |d_f(\zeta(y),\zeta(y')) - d_{f_{\mathrm{SP}}}(\zeta(y),\zeta(y'))| \\
|d_f(x,\zeta(y')) - d_{f_{\mathrm{SP}}}(\zeta(x),y')| &\leq |d_f(x,\zeta(y')) - d_f(\zeta(x),\zeta(y'))| + |d_{f_{\mathrm{SP}}}(\zeta(x),y') \\
&\quad - d_{f_{\mathrm{SP}}}(\zeta(x),\zeta(y'))| \\
&\quad + |d_f(\zeta(x),\zeta(y')) - d_{f_{\mathrm{SP}}}(\zeta(x),\zeta(y'))| \\
&\leq \omega(\delta_n) + \omega(d_{\mathrm{H},n}) + |d_f(\zeta(x),\zeta(y')) - d_{f_{\mathrm{SP}}}(\zeta(x),\zeta(y'))|
\end{aligned}$$

Since $\omega(d_{\mathrm{H},n}) \leq \omega(\delta_n)$, one has $d_{\mathrm{GH}}((G_{\delta_n}, d_{f_{\mathrm{SP}}}), (\mathcal{X}, d_f)) \leq 2\omega(\delta_n) + \max_{x,x' \in \mathbb{X}_n} |d_f(x,x') - d_{f_{\mathrm{SP}}}(x,x')|$.

Let $x, x' \in \mathbb{X}_n$. We now find upper and lower bounds for $d_{f_{\mathrm{SP}}}(x,x') - d_f(x,x')$.

Upper bound In order to upper bound $d_{f_{\mathrm{SP}}}(x,x') - d_f(x,x')$, we first show that $d_{f_{\mathrm{SP}}}(x,x')$ cannot be arbitrarily large relative to $d_f(x,x')$. Let γ be a path on $\mathcal{X}$ from x to x' achieving $d_f(x,x')$. Since $d_{\mathrm{H}}^g(\mathbb{X}_n, \mathcal{X}) \leq d_{\mathrm{H},n}$, for each $t \in [0,1]$, there exists $x_t \in \mathbb{X}_n$ such that $d_g(\gamma(t), x_t) \leq d_{\mathrm{H},n}$. Moreover, since $\mathbb{X}_n$ is finite, the set $\{x_t : t \in [0,1]\}$ can be written as $\{x_{t_1}, \ldots, x_{t_m}\}$ for some $m \in \mathbb{N}^*$, with $t_1 \leq \cdots \leq t_m$. Moreover, we claim that $\|x_{t_i} - x_{t_{i+1}}\| \leq \delta_n$, i.e., the set $\{x, x_{t_1}, \ldots, x_{t_m}, x'\}$ forms a path in G_{δ_n}. Indeed:

$$\begin{aligned}
\|x_{t_i} - x_{t_{i+1}}\| &\leq \|x_{t_i} - \gamma(t_i)\| + \|\gamma(t_i) - \gamma(t_{i+1})\| + \|\gamma(t_{i+1}) - x_{t_{i+1}}\| \\
&\leq d_g(x_{t_i}, \gamma(t_i)) + d_g(\gamma(t_i), \gamma(t_{i+1})) + d_g(\gamma(t_{i+1}), x_{t_{i+1}}) \\
&\leq 2d_{\mathrm{H},n} + d_g(\gamma(t_i), \gamma(t_{i+1}))
\end{aligned}$$

The geodesic distance $d_g(\gamma(t_i), \gamma(t_{i+1}))$ is necessarily less than $4d_{\mathrm{H},n}$, otherwise it would be possible to find a point along the geodesic, say $\gamma(\bar{t})$, such that $t_i \leq \bar{t} \leq t_{i+1}$ and $d_g(\gamma(\bar{t}), \gamma(t_i)) > 2d_{\mathrm{H},n}$ and $d_g(\gamma(\bar{t}), \gamma(t_{i+1})) > 2d_{\mathrm{H},n}$, which lead to $d_g(\gamma(\bar{t}), x_{t_i}) > d_{\mathrm{H},n}$ and $d_g(\gamma(\bar{t}), x_{t_{i+1}}) > d_{\mathrm{H},n}$, contradicting $d_{\mathrm{H}}^g(\mathbb{X}_n, \mathcal{X}) \leq d_{\mathrm{H},n}$. Hence, $\|x_{t_i} - x_{t_{i+1}}\| \leq 6d_{\mathrm{H},n} \leq \delta_n$ by assumption. Let γ' be the path from x to x' in G_{δ_n} that goes through the points $\{x, x_{t_1}, \ldots, x_{t_m}, x'\} \in \mathbb{X}_n$. We also use x_0 and x_1 to denote x and x'. Then

$$\begin{aligned}
d_{f_{\mathrm{SP}}}(x,x') &\leq \mathrm{diam}_{\mathcal{Z}}(f_{\mathrm{SP}} \circ \gamma') \\
&\leq \max\{d_{\mathcal{Z}}(f(u), f(v)) : u, v \in \{x, x_{t_1}, \ldots, x_{t_m}, x'\})\} \\
&\leq d_f(x,x') + 2 \cdot \max\{d_{\mathcal{Z}}(f(x_t), f(\gamma(t))) : t \in \{0, t_1 \ldots, t_m, 1\}\} \\
&\leq d_f(x,x') + 2\omega(d_{\mathrm{H},n}) \leq d_f(x,x') + 2\omega(\delta_n)
\end{aligned}$$

Lower bound Finally, we now show that $d_{f_{\mathrm{SP}}}(x, x')$ cannot be arbitrarily small relative to $d_f(x, x')$. Let γ be a path in G_{δ_n} achieving $d_{f_{\mathrm{SP}}}(x, x')$. Let $\gamma \cap \mathbb{X}_n = \{x_0, x_1, \ldots, x_m, x_{m+1}\}$, i.e., γ goes through the points $x_0, \ldots, x_{m+1} \in \mathbb{X}_n$ with $x_0 = x$ and $x_{m+1} = x'$. Finally, let γ' be the path from x to x' in $\mathcal{X}$ defined with $\gamma' = \gamma_m \circ \cdots \circ \gamma_0$, where γ_i is a path achieving $d_g(x_i, x_{i+1})$.

Now, we claim that $\gamma' \subseteq \bigcup_{0 \le i \le m+1} B_g(x_i, (\pi/2)\delta_n)$. Indeed, it follows from Lemma 3 in Boissonnat et al. (2019) that $\|x_i - x_{i+1}\| \le \delta_n \le 2 \cdot \mathrm{rch}(\mathcal{X}) \Rightarrow d_g(x_i, x_{i+1}) \le 2 \cdot \mathrm{rch}(\mathcal{X}) \cdot \arcsin\left(\frac{\|x_i - x_{i+1}\|}{2 \cdot \mathrm{rch}(\mathcal{X})}\right) \le (\pi/2)\|x_i - x_{i+1}\| \le (\pi/2)\delta_n$. Then, one has

$$
\begin{aligned}
d_f(x, x') &\le \mathrm{diam}_{\mathcal{Z}}(f \circ \gamma') \\
&\le \mathrm{diam}_{\mathcal{Z}}(f(\bigcup\nolimits_{0 \le i \le m+1} B_g(x_i, (\pi/2)\delta_n))) \\
&= \sup\{d_{\mathcal{Z}}(f(u), f(v)) \,:\, u, v \in \bigcup\nolimits_{0 \le i \le m+1} B_g(x_i, (\pi/2)\delta_n)\} \\
&\le \sup\{d_{\mathcal{Z}}(f(u), f(v)) \,:\, u, v \in \{x_0, \ldots, x_{m+1}\}\} \\
&\quad + 2 \cdot \max_i \mathrm{diam}_{\mathcal{Z}}(f(B_g(x_i, (\pi/2)\delta_n))) \\
&\le d_{f_{\mathrm{SP}}}(x, x') + 2\omega((\pi/2)\delta_n)
\end{aligned}
$$

We can finally conclude: $d_{\mathrm{GH}}((G_{\delta_n}, d_{f_{\mathrm{SP}}}), (\mathcal{X}, d_f)) \le 2\omega(\delta_n) + \max_{x,x' \in \mathbb{X}_n} |d_f(x, x') - d_{f_{\mathrm{SP}}}(x, x')| \le 2\omega(\delta_n) + 2\omega((\pi/2)\delta_n) \le 4\omega(2\delta_n)$. □

We are now ready to prove Theorem 3.5.

Proof **Theorem** 3.5. We first decompose the objective into three terms:

$$
\begin{aligned}
&\mathbb{E}\left[d_{\mathrm{GH}}((\mathrm{M}_n, \tilde{d}_{\hat{f}_{\mathrm{SP}},\mathcal{U}}), (\mathrm{R}_f(\mathcal{X}), \tilde{d}_f))\right] \\
&\quad = \mathbb{E}\left[d_{\mathrm{GH}}((\mathrm{M}_{\hat{f}_{\mathrm{SP}},\mathcal{U}}(G_{\delta_n}), \tilde{d}_{\hat{f}_{\mathrm{SP}},\mathcal{U}}), (\mathrm{R}_f(\mathcal{X}), \tilde{d}_f))\right] \text{ by Lemma } B.1 \\
&\quad \le \mathbb{E}\left[d_{\mathrm{GH}}((\mathrm{M}_{\hat{f}_{\mathrm{SP}},\mathcal{U}}(G_{\delta_n}), \tilde{d}_{\hat{f}_{\mathrm{SP}},\mathcal{U}}), (\mathrm{R}_{\hat{f}_{\mathrm{SP}}}(G_{\delta_n}), \tilde{d}_{\hat{f}_{\mathrm{SP}}})) \cdot \mathbb{1}_\Omega\right] \qquad (11) \\
&\qquad + \mathbb{E}\left[d_{\mathrm{GH}}((\mathrm{R}_{\hat{f}_{\mathrm{SP}}}(G_{\delta_n}), \tilde{d}_{\hat{f}_{\mathrm{SP}}}), (\mathrm{R}_f(\mathcal{X}), \tilde{d}_f)) \cdot \mathbb{1}_\Omega\right] \\
&\qquad + \mathbb{P}(\Omega^c) \cdot \omega(D_{\mathcal{X}}), \qquad (12)
\end{aligned}
$$

where Ω is the event $\{d_{\mathrm{H},n} \le \delta_n/6\} \cap \{\delta_n \le 2 \cdot \mathrm{rch}(\mathcal{X})\}$, and $D_{\mathcal{X}}$ is the diameter of $\mathcal{X}$.

Let us now bound (11) and (12):

- Term (11). According to Theorem 2.11, we have

$$
\mathbb{E}\left[d_{\mathrm{GH}}((\mathrm{M}_{\hat{f}_{\mathrm{SP}},\mathcal{U}}(G_{\delta_n}), \tilde{d}_{\hat{f}_{\mathrm{SP}},\mathcal{U}}), (\mathrm{R}_{\hat{f}_{\mathrm{SP}}}(G_{\delta_n}), \tilde{d}_{\hat{f}_{\mathrm{SP}}}))\right] \le 5 \cdot \mathbb{E}\left[\mathrm{res}(\mathcal{U}, \hat{f}_{\mathrm{SP}})\right].
$$

- Term (12). According to Lemma B.3, we have:

$$
\mathbb{E}\left[d_{\mathrm{GH}}((\mathrm{R}_{\hat{f}_{\mathrm{SP}}}(G_{\delta_n}), \tilde{d}_{\hat{f}_{\mathrm{SP}}}), (\mathrm{R}_f(\mathcal{X}), \tilde{d}_f))\right] \le 4\mathbb{E}\left[\omega(2\delta_n)\right] + 2\|f_{\mathrm{SP}} - \hat{f}_{\mathrm{SP}}\|_\infty.
$$

We conclude with Lemma B.4. □

Lemma B.4 *Under assumptions* (H1), (H2) *and* (H3), *and for* Ω *defined as before, one has*

$$\mathbb{P}(\Omega^c) \cdot \omega(D_{\mathcal{X}}) + 4\mathbb{E}[\omega(2\delta_n)] \leq C\omega\left(C' \frac{\log(n)^{(2+\beta)/b}}{n^{1/b}}\right)$$

where C, C' *only depends on a, b and on the geometric parameters of* $\mathcal{X}$.

Proof The proof is borrowed from Appendix A.7 in Carrière et al. (2018), see this reference for more details on the proof. Let $K = 2 \cdot \mathrm{rch}(\mathcal{X})$. Note that by definition:

$$\mathbb{P}(\Omega^c) + 4\mathbb{E}[\omega(2\delta_n)] \leq \mathbb{P}(d_{\mathrm{H},n} > \delta_n/6) + \mathbb{P}(\delta_n > K) + 4\mathbb{E}[\omega(2\delta_n)].$$

Moreover, since P is (a, b)-standard, one has:

$$\mathbb{P}(d_{\mathrm{H}}^E(\mathbb{X}_n, \mathcal{X}) \geq u) \leq \min\left\{1, \frac{4^b}{au^b} \mathrm{e}^{-a\left(\frac{u}{2}\right)^b n}\right\} = f_{a,b}(n, u), \forall u > 0. \tag{13}$$

Let us bound each term independently.

Second term We have the following inequalities:

$$\begin{aligned}
\mathbb{P}(\delta_n > K) &= \mathbb{P}(d_{\mathrm{H}}^E(\mathbb{X}_{t(n)}, \mathbb{X}_n) > K) \\
&\leq \mathbb{P}(d_{\mathrm{H}}^E(\mathbb{X}_{t(n)}, \mathcal{X}) + d_{\mathrm{H}}^E(\mathbb{X}_n, \mathcal{X}) > K) \\
&\leq \mathbb{P}(d_{\mathrm{H}}^E(\mathbb{X}_{t(n)}, \mathcal{X}) > K/2 \cup d_{\mathrm{H}}^E(\mathbb{X}_n, \mathcal{X}) > K/2) \\
&\leq \mathbb{P}(d_{\mathrm{H}}^E(\mathbb{X}_{t(n)}, \mathcal{X}) > K/2) + \mathbb{P}(d_{\mathrm{H}}^E(\mathbb{X}_n, \mathcal{X}) > K/2) \\
&\leq f_{a,b}(t(n), K/2) + f_{a,b}(n, K/2)
\end{aligned}$$

First term(See term (B) in the proof of Proposition 13 in Carrière et al. 2018 for more details). Note that when $d_{\mathrm{H}}^E(\mathbb{X}_n, \mathcal{X}) \leq K$, it follows from Lemma 3 in Boissonnat et al. (2019) that $d_{\mathrm{H},n} \leq (\pi/2) d_{\mathrm{H}}^E(\mathbb{X}_n, \mathcal{X})$. Thus, we have the following inequalities:

$$\begin{aligned}
\mathbb{P}(d_{\mathrm{H},n} > \delta_n/6) &\leq \mathbb{P}(d_{\mathrm{H},n} > \delta_n/6 \cap d_{\mathrm{H}}^E(\mathbb{X}_n, \mathcal{X}) \leq K) + \mathbb{P}(d_{\mathrm{H}}^E(\mathbb{X}_n, \mathcal{X}) > K) \\
&\leq \mathbb{P}(d_{\mathrm{H}}^E(\mathbb{X}_n, \mathcal{X}) > \delta_n/(3\pi) \cap d_{\mathrm{H}}^E(\mathbb{X}_n, \mathcal{X}) \leq K) + \mathbb{P}(d_{\mathrm{H}}^E(\mathbb{X}_n, \mathcal{X}) > K) \\
&\leq \mathbb{P}(d_{\mathrm{H}}^E(\mathbb{X}_n, \mathcal{X}) > \delta_n/(3\pi)) + \mathbb{P}(d_{\mathrm{H}}^E(\mathbb{X}_n, \mathcal{X}) > K) \\
&\leq \frac{2^{b-1}}{n\log(n)} + f_{a,b}(n, K) \text{ for } n \text{ large enough},
\end{aligned}$$

since it is known that, given a constant $C > 0$, the probability $\mathbb{P}(d_{\mathrm{H}}^E(\mathbb{X}_n, \mathcal{X}) > C\delta_n)$ is always upper bounded by $\frac{2^{b-1}}{n\log(n)}$ for n large enough (with the minimal required value for n increasing with the constant C and the ambient dimension).

Third term(See term (A) in the proof of Proposition 13 in Carrière et al. (2018) for more details). This is the dominating term. Let $\bar{D} = \omega(2D_{\mathcal{X}})$. Then, we have:

$$\begin{aligned}\mathbb{E}[\omega(2\delta_n)] &= \int_0^{\bar{D}} \mathbb{P}(\omega(2\delta_n) \geq \alpha)\mathrm{d}\alpha \\ &\leq \int_0^{\bar{D}} \mathbb{P}\left(d_{\mathrm{H}}^E(\mathbb{X}_n, \mathcal{X}) \geq \frac{1}{4}\omega^{-1}(\alpha)\right)\mathrm{d}\alpha \\ &\quad + \int_0^{\bar{D}} \mathbb{P}\left(d_{\mathrm{H}}^E(\mathbb{X}_{t(n)}, \mathcal{X}) \geq \frac{1}{4}\omega^{-1}(\alpha)\right)\mathrm{d}\alpha \\ &\leq C''\omega\left[\left(C'\frac{\log(t(n))}{t(n)}\right)^{1/b}\right],\end{aligned}$$

where the constants C', C'' depend on a, b. □

B.2 Proof of Theorem 3.8

In this section, we have $\mathcal{Z} = \mathbb{R}^p$. The notation $\|\cdot\|$ is the Euclidean norm either in $\mathbb{R}^D$ or in $\mathbb{R}^p$. The constant C may change from line to line.

B.2.1 Preliminary results

We consider an optimal k-points $\hat{t} := t(P_n^{\hat{f}})$ for the measure $P_n^{\hat{f}}$. Let us introduce the distance function $d_{\hat{t}}$ to a k-points $\hat{t}$ of $(\mathbb{R}^p)^k$: for any $z \in \mathbb{R}^p$,

$$d_{\hat{t}}(z) = \min_{j=1,\ldots,k} \|z - \hat{t}_j\|.$$

We also introduce the random variable

$$\Delta = \sup_{i=1,\ldots,n} d_{\hat{t}}(f(X_i)).$$

Let $\widehat{\mathcal{U}}^\epsilon = \{\hat{U}_j^\epsilon\}_{j=1,\ldots,k}$ be the ϵ-thickening of the Voronoi partition associated to $\hat{t}$. We start with the following lemma:

Lemma B.5 *Under assumptions* (H1)*,* (H3) *and* (H4)*,*

$$\frac{1}{2}\mathrm{res}(\widehat{\mathcal{U}}^\epsilon, \hat{f}_{\mathrm{SP}}) \leq \Delta + 3\|(f - \hat{f})|_{\mathbb{X}_n}\|_\infty + \omega(\delta_n) + \varepsilon.$$

Proof Let $j \in \{1, \ldots, k\}$ and $z' \in \hat{U}_j^\epsilon \cap \mathrm{im}(\hat{f}_{\mathrm{SP}})$. There exists $z \in \mathrm{im}(\hat{f}_{\mathrm{SP}})$ such that $\|z - z'\| \leq \varepsilon$, and z belongs to the j-th Voronoi cell associated to $\hat{t}$, i.e., $d_{\hat{t}}(z) = \|z - \hat{t}_j\|$.

Let $x \in G_{\delta_n}$ such that $z = \hat{f}_{\mathrm{SP}}(x)$. The point x belongs to an edge $[X_{i_1}, X_{i_2}]$ of G_{δ_n}. We have

$$\begin{aligned}
\|z' - \hat{t}_j\| &\leq \|z - \hat{t}_j\| + \|z' - z\| \\
&= d_{\hat{t}}(\hat{f}_{\mathrm{SP}}(x)) + \|z' - z\| \\
&\leq d_{\hat{t}}(\hat{f}_{\mathrm{SP}}(x)) + \varepsilon \\
&\leq d_{\hat{t}}(\hat{f}(X_{i_1})) + \|\hat{f}_{\mathrm{SP}}(X_{i_1}) - \hat{f}_{\mathrm{SP}}(x)\| + \varepsilon \\
&\leq d_{\hat{t}}(f(X_{i_1})) + 3\|(f - \hat{f})|_{\mathbb{X}_n}\|_\infty + \|f_{\mathrm{SP}}(X_{i_1}) - f_{\mathrm{SP}}(x)\| + \varepsilon \\
&\leq \Delta + 3\|(f - \hat{f})|_{\mathbb{X}_n}\|_\infty + \omega(\delta_n) + \varepsilon,
\end{aligned}$$

where we have used the fact that $d_{\hat{t}}$ is 1-Lipschitz for the third and fourth inequalities, the fact that $\mathcal{Z} = \mathbb{R}^p$ and thus $\|f_{\mathrm{SP}} - \hat{f}_{\mathrm{SP}}\|_\infty \leq \|(f - \hat{f})|_{\mathbb{X}_n}\|_\infty$ (see remark under Theorem 3.5) for the fourth inequality, and the fact that $\|x - X_{i_1}\| \leq d_{\mathcal{X}}(x, X_{i_1}) \leq \delta_n$ for the last inequality. The result follows. □

In the following, we use standard notation in the field of empirical processes: for some integrable function h with respect to some measure Q, let $Qh = \int h(x) dQ(x)$. Let P^f be the push forward measure of P by f.

Lemma B.6 *Under assumptions* (H2) *and* (H5)*, the following inequality holds conditionally to* $\mathbb{X}_n$*:*

$$\Delta \leq C \left(P^f d_{\hat{t}}^2\right)^{\frac{\gamma}{b+2\gamma}} \vee \left(P^f d_{\hat{t}}^2\right)^{\frac{1}{2}}$$

where the constant C *only depends on* a*,* b*,* c *and* γ*.*

Proof Let $z^* \in \mathbb{Z}_n$ such that $\Delta = d_{\hat{t}}(z^*)$. The function $z \in \mathcal{Z} \mapsto d_{\hat{t}}(z)$ is 1-Lipschitz, thus for any $z \in \mathcal{Z}$ such that $\|z - z^*\| \leq \frac{\Delta}{2}$, we have

$$\left|d_{\hat{t}}(z^*) - d_{\hat{t}}(z)\right| \leq \frac{\Delta}{2}.$$

This gives the inclusion

$$B\left(z^*, \frac{\Delta}{2}\right) \subseteq \left\{z \in \mathcal{Z} \,:\, d_{\hat{t}}(z) \geq \frac{\Delta}{2}\right\}.$$

Then, using Markov's inequality on P^f, we obtain

$$\begin{aligned}
P^f d_{\hat{t}}^2 &\geq \frac{\Delta^2}{4} P^f\left(\left\{z \in \mathcal{Z} \,:\, d_{\hat{t}}(z) \geq \frac{\Delta}{2}\right\}\right) \\
&\geq a\frac{\Delta^2}{4}\left[\omega^{-1}\left(\frac{\Delta}{2}\right)\right]^b \wedge \frac{\Delta^2}{4}
\end{aligned}$$

$$\geq a \frac{1}{2^b 4 c^{b/\gamma}} \Delta^{\frac{2\gamma+b}{\gamma}} \wedge \frac{\Delta^2}{4},$$

where we have used Lemma B.7 for the second inequality and (H5) for the third inequality. □

Lemma B.7 *Under assumptions* (H1), (H2) *and* (H3), *for any* $r \geq 0$ *and any* $z \in \mathrm{im}(f)$, *the push forward distribution* P^f *satisfies the inequality*

$$P^f(B(z,r)) \geq a \left[\omega^{-1}(r)\right]^b.$$

Proof For any $r \geq 0$ and any $z = f(x) \in \mathrm{im}(f)$, by definition of the push forward measure P^f,

$$\begin{aligned}
\int_{\mathcal{Z}} \mathbb{1}_{B(f(x),r)}(z') dP^f(z') &\geq \int_{\mathcal{X}} \mathbb{1}_{B(f(x),r)}(f(x')) dP(x') \\
&\geq \int_{B(x,\omega^{-1}(r))} \mathbb{1}_{B(f(x),r)}(f(x')) dP(x') \\
&\geq P\left(B(x, \omega^{-1}(r)\right) \\
&\geq a\left(\omega^{-1}(r)\right)^b.
\end{aligned}$$

where we have used for the second inequality the fact that $\omega(\omega^{-1}(u)) = u$, because ω is continuous. □

Let $t^\star = t(P^f)$ be an optimal k points for the measure P^f.

Lemma B.8 *Under assumptions* (H1), (H2) *and* (H5),

$$P^f d_{t^\star}^2 \leq C k^{-\frac{2\gamma}{b}}$$

where C *only depends on* a, b, c *and* γ.

Proof From Lemma B.7, it can be easily derived that the δ-covering number of the support of P^f is upper bounded by $C\delta^{-b/\gamma}$ where C only depends on a, c, b and γ (see for instance the proof of Lemma 10 in Chazal et al. (2015)). In other words, the minimum radius $\bar{\delta}$ to cover the support of P^f with k balls is upper bounded by $Ck^{-\gamma/b}$. There exists a family of k balls of radius $\bar{\delta}$: $B(\bar{t}_{j_1}, \bar{\delta}), \ldots, B(\bar{t}_{j_k}, \bar{\delta})$ which is a cover of the support of P^f. We also define the function $\bar{j} : z \in \mathcal{Z} \mapsto \{1, \ldots, k\}$ which returns the index j of the (or one of the) closest center $\bar{t}_j$ to any point z of the support of P^f. Consequently,

$$\begin{aligned}
P^f d_{t^\star}^2 &\leq P^f d_{\bar{t}}^2 \\
&\leq \mathbb{E}\left(\|Z - \bar{t}_{\bar{j}(Z)}\|^2\right)
\end{aligned}$$

$$\leq \mathbb{E}\left[\mathbb{E}\left(\|Z - \bar{t}_{\bar{j}(Z)}\|^2 \mathbb{1}_{\|Z-\bar{t}_{\bar{j}(Z)}\|\leq\bar{\delta}} \middle| \bar{j}(Z)\right)\right] \tag{14}$$

Conditionally to $\bar{j}(Z) = j$, one has

$$\begin{aligned}\mathbb{E}\left(\|Z - \bar{t}_j\|^2 \mathbb{1}_{\|Z-\bar{t}_j\|\leq\bar{\delta}}\right) &= \int_0^{\bar{\delta}^2} P\left(\|Z - \bar{t}_j\|^2 > u\right) du \\ &= \int_0^{\bar{\delta}^2} \left\{1 - P\left(\|Z - \bar{t}_j\|^2 \leq u\right)\right\} du \\ &\leq Ck^{-\frac{2\gamma}{b}}.\end{aligned}$$

We conclude by integrating this bound in inequality (14). □

B.2.2 Main part of the proof of Theorem 3.8

From Lemmas B.4, B.5, and B.6, and by Jensen's inequality, we find that an upper bound is

$$C\left\{\left[\mathbb{E}\left(P^f d_{\hat{t}}^2\right)\right]^{\frac{\gamma}{b+2\gamma}} + \left[\mathbb{E}\left(P^f d_{\hat{t}}^2\right)\right]^{\frac{1}{2}} + \mathbb{E}\|(f - \hat{f})|_{\mathbb{X}_n}\|_\infty + \omega\left(\frac{\log(n)^{(2+\beta)/b}}{n^{1/b}}\right)\right\} + 10\varepsilon \tag{15}$$

where C only depends on a, b, c, γ and on the geometric parameters of $\mathcal{X}$. Next we need to upper bound the expectation of $P^f d_{\hat{t}}^2$. Let P_n^f be the push forward of P_n by f, that is the empirical distribution corresponding to the Z_i's. We start with the following standard decomposition:

$$\begin{aligned}0 \leq P^f d_{\hat{t}}^2 - P^f d_{t^\star}^2 &= (P^f - P_n^f) d_{\hat{t}}^2 + P_n^f d_{\hat{t}}^2 - P^f d_{t^\star}^2 \\ &\leq (P^f - P_n^f) d_{\hat{t}}^2 + (P_n^f - P^f) d_{t^\star}^2\end{aligned}$$

where $t^\star = t(P^f)$ is an optimal k points for the measure P^f. Note that $t^\star \in (B(0, \|f\|_\infty))^k$ and that $\hat{t} \in (B(0, \|f\|_\infty))^k$ almost surely. Thus,

$$\mathbb{E}(P^f d_{\hat{t}}^2) \leq P^f d_{t^\star}^2 + 2\mathbb{E} \sup_{t\in(B(0,\|f\|_\infty))^k} \left|(P^f - P_n^f) d_t^2\right| \tag{16}$$

where the expectation is under the distribution of $\mathbb{Z}_n$.

Proposition B.9 *The following inequality holds:*

$$\mathbb{E} \sup_{t\in(B(0,\|f\|_\infty))^k} \left|(P^f - P_n^f) d_t^2\right| \leq \frac{C\|f\|_\infty^2}{\sqrt{n}} \sqrt{k(p+2)}$$

where C is an absolute constant.

Proof We introduce the functional spaces

$$\mathcal{G}_1 = \left\{ z \mapsto \|z - t_1\|^2 \mathbb{1}_{B(0,\|f\|_\infty)}(z) \,:\, t_1 \in B\left(0, \|f\|_\infty\right) \right\}$$

and

$$\begin{aligned}\mathcal{G} &= \left\{ z \mapsto d_t^2(z) \mathbb{1}_{B(0,\|f\|_\infty)}(z) \,:\, t \in \left(B\left(0, \|f\|_\infty\right)\right)^k \right\} \\ &= \left\{ z \mapsto \min_{j=1,\dots,k} l_j(z) \,:\, l_j \in \mathcal{G}_1 \right\}.\end{aligned}$$

Note that $0 \leq g \leq 4\|f\|_\infty^2$ for any $g \in \mathcal{G}$. According to Theorem B.10 and Lemma B.11,

$$\begin{aligned}\mathbb{E}\left[\sup_{g\in\mathcal{G}} \left|(P^f - P_n^f)g\right|\right] &\leq 96\frac{\|f\|_\infty^2}{\sqrt{n}}\mathbb{E}\left[\int_0^{\frac{1}{2}} \sqrt{\log\left(N'_{\|\cdot\|}\left(\frac{u}{2}, \frac{(\mathcal{G}\cup -\mathcal{G})(Z_1^n)}{4\|f\|_\infty^2\sqrt{n}}\right)\right)} du\right] \\ &\leq 96\frac{\|f\|_\infty^2}{\sqrt{n}}\mathbb{E}\left[\int_0^{\frac{1}{2}} \sqrt{\log\left(2N'_{\|\cdot\|}\left(\frac{u}{2}, \frac{\mathcal{G}(Z_1^n)}{\|f\|_\infty^2\sqrt{n}}\right)\right)} du\right] \\ &\leq 96\frac{4\|f\|_\infty^2}{\sqrt{n}}\mathbb{E}\left[\int_0^{\frac{1}{2}} \sqrt{\log 2 + k\log\left(N'_{\|\cdot\|}\left(\frac{u}{2}, \frac{\mathcal{G}_1(Z_1^n)}{4\|f\|_\infty^2\sqrt{n}}\right)\right)} du\right].\end{aligned} \tag{17}$$

According to Lemma B.11,

$$N'_{\|\cdot\|}\left(\frac{u}{2}, \frac{\mathcal{G}_1(Z_1^n)}{4\|f\|_\infty^2\sqrt{n}}\right) \leq N'_{\|\cdot\|}\left(\frac{u}{4}, \mathcal{G}_2(Z_1^n)\right) N'_{\|\cdot\|}\left(\frac{u}{4}, \mathcal{G}_3(Z_1^n)\right).$$

where

$$\mathcal{G}_2 = \left\{ z \mapsto \frac{\|t_1\|^2}{4\|f\|_\infty^2\sqrt{n}} \mathbb{1}_{B(0,\|f\|_\infty)}(z) \,:\, t_1 \in B(0, \|f\|_\infty) \right\}$$

and

$$\mathcal{G}_3 = \left\{ z \mapsto \frac{\langle z, t_1\rangle}{2\|f\|_\infty^2\sqrt{n}} \mathbb{1}_{B(0,\|f\|_\infty)}(z) : t_1 \in B(0, \|f\|_\infty) \right\}.$$

Note that $\mathcal{G}_2 \subset \mathcal{G}_4 = \left\{ z \mapsto \frac{u}{\sqrt{n}} \mathbb{1}_{B(0,\|f\|_\infty)}(z) \,:\, u \in [0, 1/4] \right\}$ and thus

$$N'_{\|\cdot\|}\left(\frac{u}{4}, \mathcal{G}_2(Z_1^n)\right) \leq N'_{\|\cdot\|}\left(\frac{u}{4}, \mathcal{G}_4(Z_1^n)\right) \leq \frac{2}{\delta}.$$

Next, according to Theorem B.13 and Lemma B.14, $N'_{\|\cdot\|}\left(\frac{u}{4}, \frac{\mathcal{G}_3(Z_1^n)}{4\|f\|_\infty^2\sqrt{n}}\right) \leq \left(\frac{8}{\delta}\right)^{c_1(p+2)}$. As a consequence,

$$\log N'_{\|\cdot\|}\left(\frac{u}{2}, \frac{\mathcal{G}_1(Z_1^n)}{4\|f\|_\infty^2\sqrt{n}}\right) \leq (1 + c_1(p+2)) \log \frac{8}{\delta}$$

and we conclude with (17). □

End of the proof of Theorem 3.8. According to inequalities (15) and (16), Proposition B.9 and Lemma B.8, and using the fact that $u \mapsto u^\zeta$ is a sub additive function for $\zeta \in (0, 1)$, it follows that one has the following upper bound, up to a constant $C > 0$,

$$\begin{aligned}
&\left[k^{-\frac{2\gamma}{b}} + \sqrt{\frac{k(p+2)}{n}}\|f\|_\infty^2\right]^{\frac{\gamma}{b+2\gamma}} + \left[k^{-\frac{2\gamma}{b}} + \sqrt{\frac{k(p+2)}{n}}\|f\|_\infty^2\right]^{\frac{1}{2}} \\
&\quad + \mathbb{E}\|(f - \hat{f})|_{\mathbb{X}_n}\|_\infty + \omega\left(\frac{\log(n)^{(2+\beta)/b}}{n^{1/b}}\right) + \frac{10}{C}\varepsilon \\
&\leq k^{-\frac{2\gamma^2}{b^2+2\gamma b}} + \left(\frac{k(p+2)}{n}\right)^{\frac{\gamma}{2b+4\gamma}} + \left(\frac{k(p+2)}{n}\right)^{\frac{1}{4}} \\
&\quad + \mathbb{E}\|(f - \hat{f})|_{\mathbb{X}_n}\|_\infty + \left(\frac{\log(n)^{2+\beta}}{n}\right)^{\gamma/b} + \frac{10}{C}\varepsilon
\end{aligned}$$

where the constant C depends on $a, b, c, \gamma, \|f\|_\infty$ and on the geometric parameters of $\mathcal{X}$. For $n \geq k(p+2)$, this upper bound can be rewritten as

$$C\left[k^{-\frac{2\gamma^2}{b^2+2\gamma b}} + \left(\frac{kp}{n}\right)^{\frac{\gamma}{2b+4\gamma}} + \mathbb{E}\|(f - \hat{f})|_{\mathbb{X}_n}\|_\infty\right] + 10\varepsilon.$$

This concludes the proof of Theorem 3.8.

B.2.3 Dudley's entropy integral and tools for covering numbers

In this section, we recall several result about Dudley's entropy integral and covering numbers. Our presentation is inspired from Section B.1 in Brécheteau and Levrard (2020).

Let $\mathcal{G}$ and $\mathcal{G}'$ be two countable families of functions $g : \mathbb{R}^p \to \mathbb{R}$. The set $\mathcal{G}(Z_1^n)$ is the set

$$\mathcal{G}(Z_1^n) = \{(g(Z_1), \ldots, g(Z_n)) \,:\, g \in \mathcal{G}\}.$$

For $S \subset \mathbb{R}^p$, let $N'_{\|\cdot\|}(\delta, S)$ denotes the δ covering number of S with respect to the euclidean norm $\|\cdot\|$ in $\mathbb{R}^p$.

Let $Z_1, \ldots, Z_n$ sampled according to P, which is a distribution on $\mathbb{R}^p$, and let P_n be the corresponding empirical measure. The next result is a particular instance of the so-called Dudley's integral.

Theorem B.10 [Boucheron and Lugosi (2013), Corollary 13.2] *Assume that $\mathcal{G}$ contains the null function and that $g \leq R$ for any $g \in \mathcal{G}$. Then,*

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} |(P - P_n)g|\right] \leq 24\frac{R}{\sqrt{n}}\mathbb{E}\left[\int_0^{\frac{1}{2}} \sqrt{\log\left(N'_{\|\cdot\|}\left(\frac{u}{2}, \frac{(\mathcal{G}\cup -\mathcal{G})(Z_1^n)}{R\sqrt{n}}\right)\right)} du\right].$$

Lemma B.11 *[Brécheteau and Levrard (2020), Lemma 33] Let $\delta > 0$. Let $\mathcal{G}_{(k)} = \{\min_{j=1,\ldots,k} g_j \ : \ g_j \in \mathcal{G}\}$ and $\mathcal{G} + \mathcal{G}' = \{g + g' \ , \ g \in \mathcal{G}, g' \in \mathcal{G}'\}$. The following inequalities hold:*

- $N'_{\|\cdot\|}\left(\delta, (\mathcal{G}\cup -\mathcal{G})(z_1^n)\right) \leq 2N'_{\|\cdot\|}\left(\delta, \mathcal{G}(z_1^n)\right)$
- $N'_{\|\cdot\|}\left(\delta, \mathcal{G}_{(k)}(z_1^n)\right) \leq \left(N'_{\|\cdot\|}\left(\delta, \mathcal{G}(z_1^n)\right)\right)^k$
- $N'_{\|\cdot\|}\left(2\delta, (\mathcal{G} + \mathcal{G}')(z_1^n)\right) \leq N'_{\|\cdot\|}\left(\delta, \mathcal{G}(z_1^n)\right) N'_{\|\cdot\|}\left(\delta, \mathcal{G}'(z_1^n)\right).$

It is possible to control the covering number of a set $\mathcal{G}(z_1^n)$ by the δ-fat-dimension (also called δ-shattering dimension) of the family $\mathcal{G}$.

Definition B.12 Let $\delta > 0$.

- A set $\{z_1, \ldots, z_m\} \subset \mathbb{R}^p$ is said to be δ-shattered by $\mathcal{G}$ if there exists $(u_1, \ldots, u_m) \in \mathbb{R}^m$ such that for all $(\varepsilon_1, \ldots, \varepsilon_m) \in \{-1, +1\}^m$, there exists $g \in \mathcal{G}$ such that:

$$\forall i \in \{1, \ldots, m\}, \ \varepsilon_i(g(z_i) - u_i) \geq \delta.$$

- The δ fat-dimension of $\mathcal{G}$, $\mathrm{fat}_\delta(\mathcal{G})$, is the size of the largest set in $\mathbb{R}^p$ that is δ-shattered by $\mathcal{G}$.

Theorem B.13 [Mendelson and Vershynin (2003), Theorem 1] *Assume that class of functions $\mathcal{G}$ is bounded by 1. There exists absolute constants c_1 and c_2 such that for all $z_1^n \in (\mathbb{R}^p)^n$ and all $\delta \in (0, 1)$,*

$$N'_{\|\cdot\|}\left(\delta, \frac{1}{\sqrt{n}}\mathcal{G}(z_1^n)\right) \leq \left(\frac{2}{\delta}\right)^{c_1 \mathrm{fat}_{c_2\gamma}(\mathcal{G})}.$$

Lemma B.14 *[Brécheteau and Levrard (2020), Lemma 37] Let $R > 0$ and $\mathcal{H} = \{z \mapsto \frac{1}{R}\mathbb{1}_{B(0,R)}(z)\langle z, v\rangle \ : \ v \in S(0,1)\}$ where $S(0,1)$ is the unit sphere of $\mathbb{R}^p$. Then, for any $\delta > 0$,*

$$\mathrm{fat}_\delta(\mathcal{H}) \leq p + 2.$$

References

Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J.: A public domain dataset for human activity recognition using smartphones. In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (2013)

Dmitri, B., Yuri, B., Sergei, I.: A Course in Metric Geometry. American Mathematical Society, Providence (2001)

Boucheron, S., Bousquet, O., Lugosi, G.: Theory of classification: a survey of some recent advances. ESAIM Probab. Stat. **9**, 323–375 (2005)

Brown, A., Bobrowski, O., Munch, E., Wang, B.: Probabilistic convergence and stability of random Mapper graphs. In: CoRR (2019). arXiv:1909.03488

Blanchard, G., Bousquet, O., Zwald, L.: Statistical properties of kernel principal component analysis. Mach. Learn. **66**(2–3), 259–294 (2007)

Brüel-Gabrielsson, R., Carlsson, G.: Exposition and interpretation of the topology of neural networks. In: CoRR (2018). arXiv:1810.03234

Bauer, U., Ge, X., Wang, Y.: Measuring distance between Reeb graphs. In: 30th Annual Symposium on Computational Geometry (SoCG 2014), pp 464–473. Association for Computing Machinery (2014)

Billingsley, P.: Convergence of Probability Measures. Wiley, Hoboken (2013)

Brécheteau, C., Levrard, C.: A k-points-based distance for robust geometric inference. Bernoulli **26**(4), 3017–3050 (2020)

Boucheron, S., Lugosi, G., Pascal, M.: Concentration Inequalities?: A Nonasymptotic Theory of Independence. Oxford University Press, Oxford (2013)

Boissonnat, J.-D., Lieutier, A., Wintraecken, M.: The reach, metric distortion, geodesic convexity and the variation of tangent spaces. J. Appl. Comput. Topol. **3**(1–2), 29–58 (2019)

Biau, G., Mas, A.: PCA-Kernel estimation. Stat. Risk Model. **29**(1), 19–46 (2012)

Chazal, F., Glisse, M., Labruère, C., Michel, B.: Convergence rates for persistence diagram estimation in topological data analysis. J. Mach. Learn. Res. **16**(110), 3603–3635 (2015)

Chazal, F., Michel, B.: An introduction to topological data analysis: fundamental and practical aspects for data scientists (2017). arXiv preprint arXiv:1710.04019

Carrière, M., Michel, B., Oudot, S.: Statistical analysis and parameter selection for Mapper. J. Mach. Learn. Res. **19**(12), 1–39 (2018)

Carrière, M., Oudot, S.: Structure and stability of the one-dimensional Mapper. Found. Comput. Math. **18**(6), 1333–1396 (2017)

Carrière, M., Rabadán, R.: Topological data analysis of single-cell Hi-C contact maps. In: The Abel Symposium 2018, vol. 15. Springer-Verlag (2018)

DeVore, R., Lorentz, G.: Constructive Approximation, vol. 303. Springer, Berlin (1993)

Dey, T., Mémoli, F., Wang, Y.: Topological analysis of nerves, Reeb spaces, mappers, and multiscale mappers. In: 33rd International Symposium on Computational Geometry (SoCG 2017), vol. 77, pp 36:1–36:16. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik (2017)

Dey, T., Mémoli, F., Wang, Y.: (2017) Topological analysis of nerves, reeb spaces, mappers, and multiscale mappers. In: CoRR. arXiv:1703.07387

de Silva, V., Munch, E., Patel, A.: Categorified Reeb graphs. Discrete Comput. Geom. **55**(4), 854–906 (2016)

Efromovich, S.: Conditional density estimation in a regression setting. Ann. Stat. **35**(6), 2504–2535 (2007)

Ge, X., Safa, I., Belkin, M., Wang, Y.: Data skeletonization via Reeb graphs. In: Advances in Neural Information Processing Systems 24 (NeurIPS 2011), pp. 837–845. Curran Associates, Inc (2011)

Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, Berlin (2003)

Jeitziner, R., Carrière, M., Rougemont, J., Oudot, S., Hess, K., Brisken, C.: Two-tier Mapper, an unbiased topology-based clustering method for enhanced global gene expression analysis. Bioinformatics **35**(18), 3339–3347 (2019)

Murtagh, F., Contreras, P.: Algorithms for hierarchical clustering: an overview. Wiley Interdiscip Rev, Data Mining Knowl. Discov. **2**(1), 86–97 (2012)

Ma, Y., Fu, Y.: Manifold Learning Theory and Applications. CRC Press, Boca Raton (2011)

Mendelson, S., Vershynin, R.: Entropy and the combinatorial dimension. Invent. Math. **152**(1), 37–55 (2003)

Munch, E., Wang, B.: Convergence between categorical representations of Reeb space and Mapper. In: 32nd International Symposium on Computational Geometry (SoCG 2016), vol. 51, pp. 53:1–53:16. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik (2016)

Nicolau, M., Levine, A., Carlsson, G.: Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proc. Natl. Acad. Sci. U.S.A. **108**(17), 7265–7270 (2011)

Naitzat, G., Lokare, N., Silva, J., Kaynar-Kabul, I.: M-Boost: profiling and refining deep neural networks with topological data analysis. In: KDD Workshop on Interactive Data Exploration and Analytics (2018)

Rizvi, A., Cámara, P., Kandror, E., Roberts, T., Schieren, I., Maniatis, T., Rabadán, R.: Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. Nat. Biotechnol. **35**, 551–560 (2017)

Reeb, G.: Sur les points singuliers d'une forme de Pfaff complètement intégrable ou d'une fonction numérique. Comptes Rendus de l'Académie des Sciences de Paris **222**, 847–849 (1946)

Reiß, M., Wahl, M., et al.: Nonasymptotic upper bounds for the reconstruction error of PCA. Ann. Stat. **48**(2), 1098–1123 (2020)

Singh, G., Mémoli, F., Carlsson, G.: Topological methods for the analysis of high dimensional data sets and 3D object recognition. In: 4th Eurographics Symposium on Point-Based Graphics (SPBG 2007), pp 91–100. The Eurographics Association (2007)

Shawe-Taylor, J., Williams, C.K.I., Cristianini, N., Kandola, J.: On the eigenspectrum of the gram matrix and the generalization error of kernel-pca. IEEE Trans. Inf. Theory **51**(7), 2510–2522 (2005)