Convergence rates for persistence diagram estimation in Topological Data Analysis

Frédéric Chazal FREDERIC.CHAZAL@INRIA.FR
INRIA Saclay Île-de-France, Palaiseau, France

Marc Glisse MARC.GLISSE@INRIA.FR
INRIA Saclay Île-de-France, Palaiseau, France

Catherine Labruère CLABRUER@U-BOURGOGNE.FR
Institut de Mathématiques de Bourgogne, France

Bertrand Michel BERTRAND.MICHEL@UPMC.FR
LSTA, Université Pierre et Marie Curie, Paris

Abstract

Computational topology has recently seen an important development toward data analysis, giving birth to Topological Data Analysis. Persistent homology appears as a fundamental tool in this field. We show that the use of persistent homology can be naturally considered in general statistical frameworks . We establish convergence rates of persistence diagrams associated to data randomly sampled from any compact metric space to a well defined limit diagram encoding the topological features of the support of the measure from which the data have been sampled. Our approach relies on a recent and deep stability result for persistence that allows to relate our problem to support estimation problems (with respect to the Gromov-Hausdorff distance). Some numerical experiments are performed in various contexts to illustrate our results.

1. Introduction

Motivations. During the last decades, the wide availability of measurement devices and simulation tools has led to an explosion in the amount of available data in almost all domains of science, industry, economy and even everyday life. Often these data come as point clouds sampled in possibly high (or infinite) dimensional spaces. They are usually not uniformly distributed in the embedding space but carry some geometric structure (manifold or more general stratified space) which reflects important properties of the "systems" from which they have been generated. Moreover, in many cases data are not embedded in Euclidean spaces and come as (finite) sets of points with pairwise distance information. This often happens, e.g. with social network or sensor network data where each sensor may not know its own position, but may evaluate its distance to the other sensors using the strength of the signal received from them. In such cases, data are given as matrices of pairwise distances between the observations, i.e. as (discrete) metric spaces. Again, although they come as abstract spaces, these data often carry specific topological and geometric structures.

A large amount of research has been done on dimensionality reduction, manifold learning and geometric inference for data embedded in Euclidean spaces and assumed to be concentrated around submanifolds; see for instance (Wang, 2012). However, the assumption that data lies on a manifold may fail in many applications. In addition, the strategy of representing data by points in Euclidean spaces may introduce large metric distortions as the data may lie in highly curved spaces. With the emergence of new geometric inference and algebraic topology tools, computational topology (Edelsbrunner & Harer, 2010) has recently seen an important development toward data analysis, giving birth to the field of Topological Data Analysis (TDA) (Carlsson, 2009)-(https://sites.google.com/site/nips2012topology/) whose aim is to infer relevant, multiscale, qualitative and, quantitative topological structures directly from the data. Topological persistence, more precisely *persistent homology* appears as a fundamental tool for TDA.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

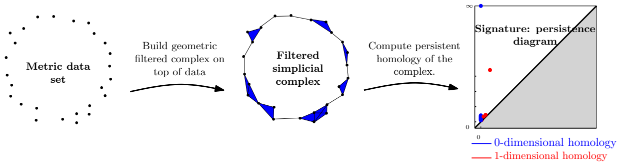
Metric data set
Build geometric filtered complex on top of data
Filtered simplicial complex
Compute persistent homology of the complex.
Signature: persistence diagram
∞
0
0
0-dimensional homology
1-dimensional homology

Figure 1. A classical pipeline for persistence in TDA.

Roughly, *homology* (with coefficient in a field such as, e.g., $\mathbb{Z}/2\mathbb{Z}$) associates to any topological space $\mathbb{M}$, a family of vector spaces (the so-called homology groups) $H_k(\mathbb{M})$, $k = 0, 1, \ldots$, each of them encoding topological features of $\mathbb{M}$. The k^{th} *Betti number* of $\mathbb{M}$, denoted β_k, is the dimension of $H_k(\mathbb{M})$ and measures the number of k-dimensional features of $\mathbb{M}$: for example, β_0 is the number of connected components of $\mathbb{M}$, β_1 the number of independent cycles or "tunnels", β_2 the number of "voids", etc. (see (Hatcher, 2001)). Persistent homology provides a framework (Edelsbrunner et al., 2002)-(Zomorodian & Carlsson, 2005)-(Chazal et al., 2012) and efficient algorithms to encode the evolution of the homology of families of nested topological spaces indexed by a set of real numbers that may often be seen as scales, such as the sublevel sets of a function, the union of growing balls, etc. The obtained multiscale topological information is then represented in a simple way as a barcode or persistence diagram; see Figure 4 and Section 2.
In TDA, persistent homology has found applications in many fields, including neuroscience (Singh et al., 2008), bioinformatics (Kasson et al., 2007), shape classification (Chazal et al., 2009b), clustering (Chazal et al., 2011b) and sensor networks (De Silva & Ghrist, 2007). It is usually computed for a *filtered simplicial complex* built on top of the available data, i.e. a nested family of simplicial complexes whose vertex set is the data set (see Section 2). The obtained persistence diagrams are then used as "topological signatures" to exhibit and compare the topological structure underlying the data; see Figure 1. The relevance of this approach relies on stability results ensuring that close data sets, with respect to the Hausdorff or Gromov-Hausdorff distance, have close persistence diagrams (Cohen-Steiner et al., 2005)-(Chazal et al., 2009a)-(Chazal et al., 2012)-(Chazal et al., 2013a). However these results are not statistical and thus only provide heuristic or exploratory uses in data analysis. Moreover,

The goal of this paper is to show that, thanks to recent results (Chazal et al., 2012)-(Chazal et al., 2013a) that allow to consider persistence diagrams associated to infinite spaces, the use of persistent homology in TDA can be naturally considered in general statistical frameworks and persistence diagrams can be used as statistics with interesting convergence properties.

Contribution. In this paper we assume that the available data is the realization of a probability distribution supported on an unknown compact metric space. We consider the persistent homology of different filtered simplicial complexes built on top of the data. We study, with a minimax approach, the rate of convergence of the associated persistence diagrams to some well-defined persistence diagram associated to the support of the probability distribution. More precisely, we assume that we observe a set of n points $\widehat{\mathbb{X}}_n = \{X_1 \ldots, X_n\}$ in a metric space $(\mathbb{M}, \rho)$, drawn i.i.d. from some unknown measure μ whose support is a compact set denoted $\mathbb{X}_\mu \subseteq \mathbb{M}$. We also assume that μ satisfies the so-called (a, b)-*standard assumption* for some constants $a, b > 0$: for any $x \in \mathbb{X}_\mu$ and any $r > 0$, $\mu(B(x, r)) \geqslant \min(ar^b, 1)$. We then consider the persistent homology of some filtered simplicial complexes $\mathrm{Filt}(\mathbb{X}_\mu)$ and $\mathrm{Filt}(\widehat{\mathbb{X}})$ built on top of $\mathbb{X}_\mu$ and $\widehat{\mathbb{X}}_n$ respectively and we establish convergence rates to zero of the *bottleneck distance* between their persistence diagrams. The definition of this distance is given in the second section. We illustrate our results when $\mathbb{M} = \mathbb{R}^d$ and also discuss the case where the data are corrupted by an additional Gaussian noise.

Our approach relies on the general theory of persistence modules and our results follow from two recently proven properties of persistence diagrams (Chazal et al., 2013a)-(Chazal et al., 2009a)-(Chazal et al., 2012). First, as $\mathbb{X}_\mu$ can be any compact metric space (possibly infinite), the filtered complex $\mathrm{Filt}(\mathbb{X}_\mu)$ is usually not finite or even countable and the existence of its persistence diagram cannot be established from the "classical" persistence theory (Zomorodian & Carlsson, 2005)-(Edelsbrunner et al., 2002). To overcome this issue we use the general persistence framework introduced in (Chazal et al., 2009a)-(Chazal et al., 2012). Notice that although this framework is rather abstract and theoretical, it does not have any practical drawback as only persistence diagrams of complexes built on top of finite data are computed. Second, a fundamental property of the persistence diagrams we are considering is their stability proven in (Chazal et al., 2013a) that establishes a strong connection between the persistence estimation problem and more classical support estimation problems.

Related works. Although it is attracting more and more interest, the use of persistent homology in data analysis remains widely heuristic. Despite a few promising results, the statistical analysis of persistent homology is still in its infancy. One of the first statistical results about persistent homology has been given in a parametric setting, by Bubenik and Kim in (Bubenik & Kim, 2007). They show for instance that for data sampled on an hypersphere according to a von Mises-Fisher distribution, the persistence diagrams of the density can be estimated with the parametric rate $n^{-1/2}$. However assuming that both the support and the parametric family of the distribution are known are strong assumptions hardly met in practice. Closer to our approach, statistical analysis of homology has also been proposed recently in (Balakrishnan et al., 2012) when the geometric structure underlying the data is a smooth submanifold of an Euclidean space. The persistence diagram of a space contains much more information than the homology of this space. So, it is not surprising that the convergence rates for the estimation of the persistence diagram we find in this paper are much slower than the convergence rates of (Balakrishnan et al., 2012) for the estimation of the homology. Our approach is also strongly connected to manifold estimation results obtained in (Genovese et al., 2012b). Our results extend to persistent homology and allow us to deal with general compact metric spaces. Still in the manifold setting, (Balakrishnan et al., 2013) develops several methods to find confidence sets for persistence diagrams using subsampling methods and kernel estimators among other approaches.

Both (Balakrishnan et al., 2013) and our work start from the observation that persistence diagram inference is strongly connected to the better known problem of measure support estimation. As far as we know, only few results about support estimation in general metric spaces are available. For example (De Vito et al., 2012) proposes kernel methods to tackle the support estimation problem. On the other hand, a large amount of literature is available for support estimation in $\mathbb{R}^d$; see for instance the review in (Cuevas, 2009). The convergence rate of the estimator $\widehat{\mathbb{X}}_n = \{X_1, \dots X_n\}$ to the support of the measure μ with respect to the Hausdorff distance is given in (Cuevas & Rodríguez-Casal, 2004) in $\mathbb{R}^d$. Support estimation in $\mathbb{R}^d$ has also been studied under various additional assumptions such as convexity assumptions (Rodríguez-Casal, 2007) or through boundary fragments estimation (Korostelëv & Tsybakov, 1993) just to name a few. When the measure has a density with respect to the Lebesgue measure, plug-in methods based on non parametric estimators have been proposed in (Cuevas & Fraiman, 1997) and (Tsybakov, 1997).

A few different methods have also been proposed for topology estimation in non-deterministic frameworks such as those based on deconvolution (Caillerie et al., 2011)-(Niyogi et al., 2011). Several recent attempts have also been made, with completely different approaches, to study persistence diagrams from a statistical point of view, such as (Mileyko et al., 2011) that studies probability measures on the space of persistence diagrams or (Bubenik, 2012) that introduces a functional representation of persistence diagrams, the so-called persistence landscapes, allowing means and variance of persistence diagrams to be defined. Notice that our results should easily extend to persistence landscapes.

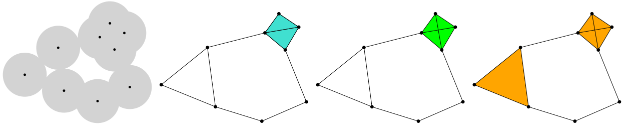

Figure 2. From left to right: the α sublevelset of the distance function to a point set $\mathbb{X}$ in $\mathbb{R}^2$, the α-complex, $\mathrm{Cech}_\alpha(\mathbb{X})$ and $\mathrm{Rips}_{2\alpha}(\mathbb{X})$. The last two include a tetrahedron.

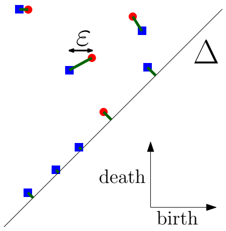

Figure 3. Two diagrams at bottleneck distance ε.

Organization of the paper. The necessary background material is presented in Section 2. The convergence results are established in Section 3 for general metric spaces and in Section 4 for measures supported in $\mathbb{R}^d$. Some numerical experiments are given in Section 5. All the proofs of the results are available in (Chazal et al., 2013b).

2. Background and notations

Measured metric spaces. Let us denote $(\mathbb{M}, \rho)$ a metric space, $\mathcal{K}(\mathbb{M})$ the set of all the compact subsets of $\mathbb{M}$, and $\mathrm{d_H}(C_1, C_2)$ the Hausdorff distance between two subsets $C_1, C_2 \in \mathcal{K}(\mathbb{M})$ (Burago et al., 2001). Note that $(\mathcal{K}(\mathbb{M}), \mathrm{d_H})$ is a metric space. Two compact metric spaces $(\mathbb{M}_1, \rho_1)$ and $(\mathbb{M}_2, \rho_2)$ are *isometric* if there exists a bijection $\Phi : \mathbb{M}_1 \to \mathbb{M}_2$ that preserves distances. One way to compare two metric spaces is to measure how far these two metric spaces are from being isometric. The corresponding distance is called the *Gromov-Hausdorff distance* (Burago et al., 2001). Intuitively, it is the infimum of their Hausdorff distance over all possible isometric embeddings of these two spaces into a common metric space. The Gromov-Hausdorff distance $\mathrm{d_{GH}}$ defines a metric on the space $\mathcal{K}$ of isometry classes of compact metric spaces (see Theorem 7.3.30 in (Burago et al., 2001)). Notice that when $\mathbb{M}_1$ and $\mathbb{M}_2$ are subspaces of a same metric space $(\mathbb{M}, \rho)$ then $\mathrm{d_{GH}}(\mathbb{M}_1, \mathbb{M}_2) \leqslant \mathrm{d_H}(\mathbb{M}_1, \mathbb{M}_2)$.

Let μ be a probability measure on $(\mathbb{M}, \rho)$ equipped with its Borel algebra. Let $\mathbb{X}_\mu$ denote the support of the measure μ, namely the smallest closed set with probability one. In the following of the paper, we will assume that $\mathbb{X}_\mu$ is compact and thus $\mathbb{X}_\mu \in \mathcal{K}(\mathbb{M})$. Also note that $(\mathbb{X}_\mu, \rho) \in \mathcal{K}$.

Geometric complexes. The geometric complexes we consider in this paper are built on top of metric spaces and come as nested families indexed by a real parameter. Topological persistence is used to infer and encode the evolution of the topology of theses families as the parameter grows. For a complete definition of these geometric filtered complexes and their use in TDA, we refer to (Chazal et al., 2013a), Section 4.2. Here, we only give a brief reminder and refer to Figure 2 for illustrations. A simplicial complex $\mathcal{C}$ is a set of simplexes (points, segments, triangles, etc) such that any face from a simplex in $\mathcal{C}$ is also in $\mathcal{C}$ and the intersection of any two simplexes of $\mathcal{C}$ is a (possibly empty) face of these simplexes. We do not assume such simplicial complexes to be finite. The complexes considered in this paper can be seen as a generalization of neighborhood graphs in dimension larger than 1.

Given a metric space $\mathbb{X}$ which will also serve as the vertex set, the *Vietoris-Rips complex* $\mathrm{Rips}_\alpha(\mathbb{X})$ is the set of simplexes $[x_0, \ldots, x_k]$ such that $d_{\mathbb{X}}(x_i, x_j) \leqslant \alpha$ for all (i, j). The *Čech complex* $\mathrm{Cech}_\alpha(\mathbb{X})$ is similarly defined as the set of simplexes $[x_0, \ldots, x_k]$ such that the $k+1$ closed balls $B(x_i, \alpha)$ have a non-empty intersection. Note that these two families of complexes only depend on the pairwise distances between the points of $\mathbb{X}$.

When $\mathbb{X}$ is embedded in $\mathbb{R}^d$, we can extend the definition of the Čech complex to the set of simplexes $[x_0, \ldots, x_k]$ such that the $k+1$ closed balls $B(x_i, \alpha)$ have a non-empty intersection in $\mathbb{R}^d$ (not just in $\mathbb{X}$). We also define the α-*complex* as the set of simplexes $[x_0, \ldots, x_k]$ such that, for some $\beta \leqslant \alpha$ that depends on the simplex, the $k+1$ closed balls $B(x_i, \beta)$ and the complement of all the other balls $B(x, \beta)$ for $x \in \mathbb{X}$ have a non-empty intersection in $\mathbb{R}^d$ (equivalently, there exists a ball of radius at most α in $\mathbb{R}^d$ such that $x_0, \ldots, x_k$ are on its boundary and the interior of the ball contains no point of $\mathbb{X}$). Those two complexes have the same homology as the union of the balls $B(x, \alpha)$ for $x \in \mathbb{X}$ (see Figure 2). The α-complex is a subcomplex of the Delaunay triangulation and thus only contains simplexes of dimension at most d. The union of the balls $B(x, \alpha)$ is also the α-sublevel set of the distance function to $\mathbb{X}$, $d(\cdot, \mathbb{X})$, and as a consequence, those filtrations provide a convenient way to study the evolution of the topology of the union of growing balls or the sublevel sets of $d(\cdot, \mathbb{X})$ (see Figure 2 and Section 5).

All these families of complexes are non-decreasing with α: for any $\alpha \leqslant \beta$, there is an inclusion of $\mathrm{Rips}_\alpha(\mathbb{X})$ in $\mathrm{Rips}_\beta(\mathbb{X})$, and similarly for the Čech, and Alpha complexes. They are called *filtrations*. In the following, the notation $\mathrm{Filt}(\mathbb{X}) := (\mathrm{Filt}_\alpha(\mathbb{X}))_{\alpha \in \mathcal{A}}$ denotes one of the filtrations defined above.

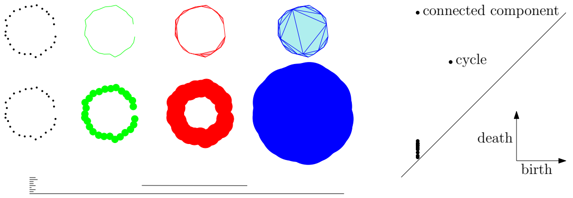

Figure 4. An α-complex filtration, the sublevelset filtration of the distance function, and their common persistence barcode.

Persistence diagrams. An extensive presentation of persistence diagrams is available in (Chazal et al., 2012). We recall a few definitions and results needed in this paper and give the intuition behind persistence.

Given a filtration, the topology of $\mathrm{Filt}_\alpha(\mathbb{X})$ changes as α increases: new connected components can appear, existing connected components can merge, cycles and cavities can appear or be filled, etc. Persistent homology tracks these changes, identifies *features* and associates an *interval* or *lifetime* (from α_{birth} to α_{death}) to them. For instance, a connected component is a feature that is born at the smallest α such that the component is present in $\mathrm{Filt}_\alpha(\mathbb{X})$, and dies when it merges with an older connected component. Intuitively, the longer a feature persists, the more relevant it is.

A feature, or more precisely its lifetime, can be represented as a segment whose extremities have abscissae α_{birth} and α_{death}; the set of these segments is called the *barcode* of $\mathrm{Filt}(\mathbb{X})$. An interval can also be represented as a point in the plane with coordinates $(\alpha_{\text{birth}}, \alpha_{\text{death}})$ (see Figure 4). The set of points (with multiplicity) representing the intervals is called the *persistence diagram* $\mathsf{dg}(\mathrm{Filt}(\mathbb{X}))$. Note that the diagram is entirely contained in the half-plane above the diagonal Δ defined by $y = x$, since death always occurs after birth. (Chazal et al., 2012) shows that this diagram is still well-defined under very weak hypotheses, and in particular $\mathsf{dg}(\mathrm{Filt}(\mathbb{X}))$ is well-defined for any compact metric space $\mathbb{X}$ (Chazal et al., 2013a). The most persistent features (supposedly the most important) are those represented by the points furthest from the diagonal in the diagram, whereas points close to the diagonal can be interpreted as (topological) noise.

The space of persistence diagrams is endowed with a metric called the *bottleneck distance* $\mathrm{d_b}$. Given two persistence diagrams, it is defined as the infimum of the δ for which we can find a matching between the diagrams, such that two points can only be matched if their distance is less than δ and all points at distance more than δ from the diagonal must be matched (see Figure 3). Note that points close to the diagonal Δ are easily ignored, which fits with their interpretation as irrelevant noise.

A fundamental property of persistence diagrams, proven in (Chazal et al., 2012), is their *stability*. If $\mathbb{X}$ and $\tilde{\mathbb{X}}$ are two compact metric spaces then one has

$$\mathrm{d_b}\left(\mathsf{dg}(\mathrm{Filt}(\mathbb{X})), \mathsf{dg}(\mathrm{Filt}(\tilde{\mathbb{X}}))\right) \leqslant 2\mathrm{d_{GH}}\left(\mathbb{X}, \tilde{\mathbb{X}}\right). \qquad (1)$$

Moreover, if $\mathbb{X}$ and $\tilde{\mathbb{X}}$ are embedded in the same space $(\mathbb{M}, \rho)$ then (1) holds for $\mathrm{d_H}$ in place of $\mathrm{d_{GH}}$. Note that these properties are metric: they do not involve any probability measure on $\mathbb{X}$ and $\tilde{\mathbb{X}}$.

3. Persistence diagrams estimation in metric spaces

Let $(\mathbb{M}, \rho)$ be a metric space. Assume that we observe n points $X_1 \dots, X_n$ in $\mathbb{M}$ drawn i.i.d. from some unknown measure μ whose support is a compact set denoted $\mathbb{X}_\mu$. The Gromov-Hausdorff distance allows us to compare $\mathbb{X}_\mu$ with compact metric spaces not necessarily embedded in $\mathbb{M}$. We thus consider $(\mathbb{X}_\mu, \rho)$ as an element of $\mathcal{K}$ (rather than an element of $\mathcal{K}(\mathbb{M})$). In the following, an *estimator* $\hat{\mathbb{X}}$ of $\mathbb{X}_\mu$ is a function of $X_1 \dots, X_n$ that takes values in $\mathcal{K}$ and is measurable for the Borel algebra induced by $\mathrm{d_{GH}}$.

Let $\mathrm{Filt}(\mathbb{X}_\mu)$ and $\mathrm{Filt}(\hat{\mathbb{X}})$ be two filtrations defined on $\mathbb{X}_\mu$ and $\hat{\mathbb{X}}$. The statistical analysis of persistence diagrams proposed below starts from the following key fact: according to (1), for any $\varepsilon > 0$:

$$\begin{aligned}\mathbb{P}\left(\mathrm{d_b}\left(\mathsf{dg}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathsf{dg}(\mathrm{Filt}(\hat{\mathbb{X}}))\right) > \varepsilon\right)& \\ \leqslant \mathbb{P}\left(\mathrm{d_{GH}}(\mathbb{X}_\mu, \hat{\mathbb{X}}) > 2\varepsilon\right)& \qquad (2)\end{aligned}$$

where the probability corresponds to the product measure $\mu^{\otimes n}$. Our strategy then consists in estimating the support $\mathbb{X}_\mu$ with respect to the $\mathrm{d_{GH}}$ distance. Note that this general strategy of estimating $\mathbb{X}_\mu$ in $\mathcal{K}$ is not only of theoretical interest. Indeed, as mentioned in the introduction, in some cases the space $\mathbb{M}$ is unknown and the observations $X_1 \dots, X_n$ are just known through their pairwise distances $\rho(X_i, X_j)$, $i, j = 1, \cdots, n$. The use of the Gromov-Hausdorff distance then allows us to consider this set of observations as an abstract metric space of cardinality n, independently of the way it is embedded in $\mathbb{M}$. This general framework includes the more standard approach consisting in estimating the support with respect to the Hausdorff distance by restraining the values of $\hat{\mathbb{X}}$ to $\mathcal{K}(\mathbb{M})$. Using equation (2), the problem of persistence diagrams estimation reduces to the better known problem of estimating the support of a measure.

Let $\hat{\mathbb{X}}_n := \{X_1, \dots, X_n\}$ be a set of independent observations endowed with the restriction of the distance ρ to this set. This finite metric space is a natural estimator of the support $\mathbb{X}_\mu$. In several contexts discussed in the following, $\hat{\mathbb{X}}_n$ shows optimal rates of convergence to $\mathbb{X}_\mu$ with respect to the Hausdorff and Gromov-Hausdorff distance. From (2) we then obtain upper bounds on the rate of convergence of $\mathrm{Filt}(\hat{\mathbb{X}}_n)$. We also obtain the corresponding lower bounds to prove optimality.

The rate of convergence of $\hat{\mathbb{X}}_n$ in Gromov-Hausdorff distance for probability measures μ satisfying an (a, d)-standard assumption is obtained using classical covering arguments - see, e.g. (Cuevas & Rodríguez-Casal, 2004; Niyogi et al., 2008). We then derive the following result for persistence diagram estimation.

Theorem 1. *Assume that the probability measure μ on $\mathbb{M}$ satisfies the (a, b)-standard assumption, then for any $\varepsilon > 0$:*

$$\begin{aligned}\mathbb{P}\left(\mathrm{d_b}\left(\mathsf{dg}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathsf{dg}(\mathrm{Filt}(\hat{\mathbb{X}}_n))\right) > \varepsilon\right)& \\ \leqslant \min\left(\frac{2^b}{a\varepsilon^b}\exp(-na\varepsilon^b), 1\right).& \qquad (3)\end{aligned}$$

Moreover,

$$\limsup_{n\to\infty}\left(\frac{n}{\log n}\right)^{1/b}\mathrm{d_b}\left(\mathsf{dg}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathsf{dg}(\mathrm{Filt}(\hat{\mathbb{X}}_n))\right) \leqslant C_1$$

almost surely, and

$$\mathbb{P}\left(\mathrm{d_b}\left(\mathsf{dg}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathsf{dg}(\mathrm{Filt}(\hat{\mathbb{X}}_n))\right) \leqslant C_2\left(\frac{\log n}{n}\right)^{1/b}\right)$$

converges to 1 *when* $n \to \infty$*, where* C_1 *and* C_2 *only depend on* a *and* b*.*

Let $\mathcal{P} = \mathcal{P}(a, b, \mathbb{M})$ be the set of all the probability measures on the metric space $(\mathbb{M}, \rho)$ satisfying the (a, b)-standard assumption on $\mathbb{M}$:

$$\begin{aligned}\mathcal{P} := \quad & \{\mu \text{ on } \mathbb{M} \mid \mathbb{X}_\mu \text{ is compact and } \forall x \in \mathbb{X}_\mu, \\ & \forall r > 0,\ \mu\left(B(x, r)\right) \geqslant \min\left(1, ar^b\right)\}.\end{aligned}$$

The next theorem gives upper and lower bounds for the rate of convergence of persistence diagrams. The upper bound is a consequence of Theorem 1, while the lower bound is established using Le Cam's lemma.

Theorem 2. *Let $(\mathbb{M}, \rho)$ be a metric space and let $a > 0$ and $b > 0$. Then:*

$$\sup_{\mu\in\mathcal{P}}\mathbb{E}\left[\mathrm{d_b}(\mathsf{dg}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathsf{dg}(\mathrm{Filt}(\hat{\mathbb{X}}_n)))\right] \leqslant C\left(\frac{\log n}{n}\right)^{1/b} \qquad (4)$$

where the constant C only depends on a and b (not on $\mathbb{M}$). Assume moreover that there exists a non isolated point x in $\mathbb{M}$ and consider any sequence $(x_n) \in (\mathbb{M}\backslash\{x\})^{\mathbb{N}}$ such that $\rho(x, x_n) \leqslant (an)^{-1/b}$. Then for any estimator $\widehat{\mathrm{dg}}_n$ of $\mathrm{dg}(\mathrm{Filt}(\mathbb{X}_\mu))$:

$$\liminf_{n\to\infty} \rho(x, x_n)^{-1} \sup_{\mu\in\mathcal{P}} \mathbb{E}\left[\mathrm{d_b}(\mathrm{dg}(\mathrm{Filt}(\mathbb{X}_\mu)), \widehat{\mathrm{dg}}_n)\right] \geqslant C'$$

where C' is an absolute constant.

Consequently, the estimator $\mathrm{dg}(\mathrm{Filt}(\widehat{\mathbb{X}}_n))$ is minimax optimal on the space $\mathcal{P}(a, b, \mathbb{M})$ up to a logarithmic term as soon as we can find a non-isolated point in $\mathbb{M}$ and a sequence (x_n) in $\mathbb{M}$ such that $\rho(x_n, x) \sim (an)^{-1/b}$. This is obviously the case for the Euclidean space $\mathbb{R}^d$.

Remark. Theorem 1 can also be used to find confidence sets for persistence diagrams. Such confidence sets depend on a and b which may be unknown and whose estimation is beyond the scope of this paper. Alternative solutions have been proposed recently in (Balakrishnan et al., 2013).

4. Persistence diagram estimation in $\mathbb{R}^k$

Persistence diagram estimation for nonsingular measures in $\mathbb{R}^k$. Paper (Singh et al., 2009) is a significant breakthrough for level set estimation through density estimation. It presents a fully data-driven procedure, in the spirit of Lepski's method, that is adaptive to unknown local density regularity and achieves a Hausdorff error control that is minimax optimal for a class of level sets with very general shapes. In this section, we study persistence diagram inference in the framework of (Singh et al., 2009). Nevertheless, we do not use the estimator of (Singh et al., 2009) for this task since we only consider here the support estimation problem (and not the more general level set issue as in (Singh et al., 2009)). Let $X_1, \ldots, X_n$ be i.i.d. observations drawn from an unknown probability measure μ having density f with respect to the Lebesgue measure on $\mathbb{R}^k$. Let $\mathbb{X}_f := \mathbb{X}_\mu$ be the support of μ and let $G_0 := \{x \in \mathbb{R}^k \mid f(x) > 0\}$. The boundary of a set G is denoted ∂G and for any $\varepsilon > 0$, $I_\varepsilon(G) := \bigcup_{x \mid B(x,\varepsilon)\subset G} B(x, \varepsilon)$ is the ε-inner of G. We denote by $\mathcal{F}(\alpha)$ the set composed of all the densities which supports are included in a fixed compact domain χ and satisfy the two following assumptions $[A]$ and $[B]$, for fixed positive constants C_a, C_b:

$[A]$: The density f is upper bounded by $f_{\max} > 0$ and there exist constants α, C_a, $\delta_a > 0$ such that for all $x \in G_0$ with $f(x) \leqslant \delta_a$, $f(x) \geqslant C_a\, d(x, \partial G_0)^\alpha$.

$[B]$: There exist constants $\varepsilon_0 > 0$ and $C_b > 0$ such that for all $\varepsilon \leqslant \varepsilon_0$, $I_\varepsilon(G_0) \neq \varnothing$ and $d(x, I_\varepsilon(G_0)) \leqslant C_b\, \varepsilon$ for all $x \in \partial G_0$.

Assumption $[A]$ describes how fast the density increases in the neighborhood of the boundary of the support: the smaller α, the easier the support may be possible to detect. Assumption $[B]$ prevents the boundary from having arbitrarily small features (as for cusps). Under assumptions $[A]$ and $[B]$, the measure μ also satisfies the standard assumption with $b = \alpha + k$. Moreover, G_0 and the support $\mathbb{X}_f$ are almost identical in the sense that $\mathrm{d_H}(G_0, \mathbb{X}_f) = 0$. As a consequence, we obtain from Theorem 2 that the estimator $\mathrm{dg}(\mathrm{Filt}(\widehat{\mathbb{X}}_n))$ converges in expectation to $\mathrm{dg}(\mathrm{Filt}(\mathbb{X}_f))$ for $\mathrm{d_b}$ with a rate upper bounded by $(\log n/n)^{1/(k+\alpha)}$. Moreover, this rate is minimax over the sets $\mathcal{F}(\alpha)$.

Persistence diagram estimation for singular measures in $\mathbb{R}^D$. We now consider the case where the support of μ is a smooth submanifold of $\mathbb{R}^D$. As far as we know, rates of convergence for manifold estimation have only been studied recently in (Genovese et al., 2012b) and (Genovese et al., 2012a) under several noise models that can be considered in the context of persistence diagram estimation. However, for the sake of simplicity, we only study here the *noiseless model* considered in (Genovese et al., 2012b). For a fixed positive integer $k < D$, for some fixed positive constants b, B, κ and for a fixed compact domain χ in $\mathbb{R}^D$, let $\mathcal{H} := \mathcal{H}(d, A, B, \kappa, \chi)$ be the set of probability measures on χ satisfying the two assumptions of (Genovese et al., 2012b): μ is in $\mathcal{H}$ if and only if $\mathbb{X}_\mu$ is a k-dimensional manifold whose reach - a measure of the regularity of $\mathbb{X}_\mu$, see Section 2 in (Genovese et al., 2012b) - is lower bounded by κ and if μ has a density g with respect to the k-dimensional volume measure on $\mathbb{X}_\mu$, such that $0 < A \leqslant \inf_{y\in\mathbb{X}_\mu} g(y) \leqslant \sup_{y\in\mathbb{X}_\mu} g(y) \leqslant B < \infty$. Under these two assumptions, μ satisfies the standard assumption with $b = k$. Thus, if we take $\hat{\mathbb{X}}_n$ for estimating the support $\mathbb{X}_\mu$ in this context, we obtain a rate of convergence upper bounded by $(\frac{\log n}{n})^{1/k}$ both for support and persistence diagram estimation. Nevertheless, this rate is not minimax optimal on the space $\mathcal{H}$, as shown by Theorem 2 in (Genovese et al., 2012b). The correct minimax rate is $n^{-2/k}$ for both estimation problems. However, the achievement of the optimal rate relies on a "theoretical" estimator that cannot be computed in practice.

Additive noise. To finish this section, we shortly discuss the additive Gaussian noise model: we now assume that the data are of the form $Y_i = X_i + \varepsilon_i$ where $X_1, \ldots X_n$ are sampled according to a measure μ as in the previous paragraph and where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. standard Gaussian random variables. We deduce from the results given in (Genovese et al., 2012b) in this context that the minimax convergence rates for the persistence diagram estimation is upper bounded by some rate of the order of $(\log n)^{-1/2}$. However, giving a tight lower bound for this problem appears to be more difficult than for the support estimation.

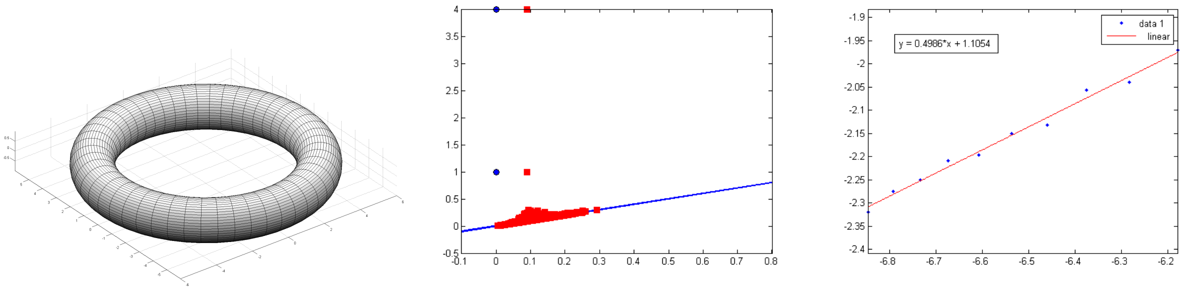

Figure 5. Left: the space $\mathbb{M}_1$. Middle: in blue the persistence diagram $\mathsf{dg}(\mathbb{M}_1)$ of the distance to $\mathbb{M}_1$ (1-dimensional homology); in red a persistence diagram of the α-filtration built on top of $n = 14000$ points randomly sampled on $\mathbb{M}_1$. Right: the log of the estimated expectation of the bottleneck distance between the estimated diagrams and $\mathsf{dg}(\mathbb{M}_1)$ as a function of $\log(\log(n)/n)$.

Table 1. Sampling parameters (where $[n : h : m]$ denotes the set $\{n, n+h, \cdots m\}$).

Space	k	n	**Geom. complex**
$\mathbb{M}_1$	100	$[4000 : 500 : 8500]$	α-complex
$\mathbb{M}_2$	20	250	Vietoris-Rips

5. Experiments

A few experiments were conducted in order to illustrate our results and the behavior of the persistence diagrams under sampling of metric spaces. We consider two different metric spaces, denoted $\mathbb{M}_1$ and $\mathbb{M}_2$ hereafter. $\mathbb{M}_1$ is the torus of revolution in $\mathbb{R}^3$ with parametric equations $x(u,v) = (5 + \cos(u))\cos(v)$, $y(u,v) = (5 + \cos(u))\sin(v)$ and $z(u,v) = \sin(u)$, $(u,v) \in [0, 2\pi]^2$ (see Figure 5, left). Its metric is the restriction of the Euclidean metric in $\mathbb{R}^3$ and it is endowed with the push forward by the parametrization of the uniform measure on the square $[0, 2\pi]^2$. $\mathbb{M}_2$ is a space of images: we used a 3D character from the SCAPE database (Anguelov et al., 2005) and considered all the images of this character from a view rotating around a fixed vertical axis. We converted these images in gray color and resized these images to $300 \times 400 = 120,000$ pixels (see Figure 6). Each is then identified with a point in $\mathbb{R}^{120,000}$ where the i^{th} coordinate is the level of gray of the i^{th} pixel. The metric space $\mathbb{M}_2$ is the obtained subset with the restriction of the Euclidean metric in $\mathbb{R}^{120,000}$. As it is parametrized by a circular set of views, it is endowed with the push forward of the uniform measure on the circle.

From each of the measured metric spaces $\mathbb{M}_1$ and $\mathbb{M}_2$, we sampled k sets of n points for different values of n from which we computed persistence diagrams for different geometric complexes (see Table 1). For $\mathbb{M}_1$ we have computed the persistence diagrams for the 1-dimensional homology of the α-complex built on top of the sampled sets. As α-complexes have the same homology as the corresponding union of balls, these persistence diagrams are the ones of the distance function to the sampled point set (Edelsbrunner, 1995). So, for each n we computed the average bottleneck distance between the obtained diagrams and the persistence diagram $\mathsf{dg}(\mathbb{M}_1)$ of the distance to $\mathbb{M}_1$ which is known exactly and represented in blue on Figure 5, middle. For each n, the average bottleneck distance between $\mathsf{dg}(\mathbb{M}_1)$ and the persistence diagrams obtained for the $k = 100$ randomly sampled sets $\mathbb{X}_n$ of size n has been used as an estimate $\hat{\mathbb{E}}$ of $\mathbb{E}\left[\mathrm{d_b}(\mathsf{dg}(\mathbb{M}_1), \mathsf{dg}(\mathrm{C}_\alpha(\widehat{\mathbb{X}}_n)))\right]$ where C_α denotes the α-complex filtration. $\log(\hat{\mathbb{E}})$ is plotted as a function of $\log(\log(n)/n)$ on Figure 5, right. As expected, since the the torus is 2-dimensional, the points are close to a line of slope $1/2$.
For $\mathbb{M}_2$, as it is embedded in a very high dimensional space, computing the α-complex is out of reach. So, we computed the persistence diagrams for the 1-dimensional homology of the Vietoris-Rips complex built on top of the sampled sets. As in that case the exact persistence diagram of the Vietoris-Rips filtration built on top of $\mathbb{M}_2$ is not known, we only computed the 1-dimensional homology persistence diagrams of the Vietoris-Rips filtrations built on top of 20 sets of 250 points each, randomly sampled on $\mathbb{M}_2$. All these diagrams have been plotted on the same Figure 6, middle. The right of Figure 6 represents a 2D embedding of one of the 250 points sampled data set using the Multidimensional Scaling algorithm (MDS). Since $\mathbb{M}_2$ is a set of images taken according to a rotating viewpoint, it carries a circular structure highlighted by the MDS embedding. The persistence diagrams that all have one point clearly off the diagonal assert the presence of a cycle in $\mathbb{M}_2$.

6. Discussion and future works

In previous works, persistent homology in TDA has been mainly considered with a non statistical approach, where persistence diagrams are used as exploratory tools to analyze the topological structure of data. The results we obtain open the door to a rigorous use of persistence diagrams in statistical frameworks.

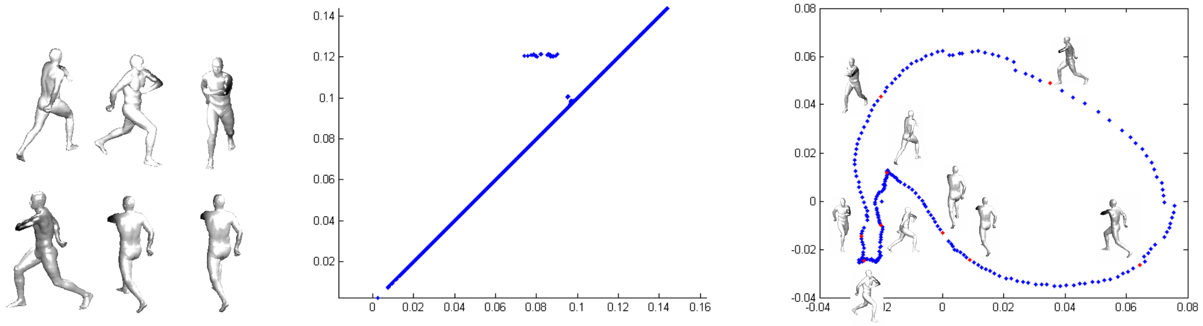

Figure 6. Left: Images sampled from $\mathbb{M}_2$. Middle: on the same figure the 1-dimensional homology persistence diagrams of the Vietoris-Rips filtration of 20 sets of 250 points sampled on $\mathbb{M}_2$. Right: the plot of the embedding of $\mathbb{M}_2$ in $\mathbb{R}^2$ using MDS.

Since our approach is based on very general recent stability results in persistence theory, it can be adapted to other frameworks. For example, building on ideas developed in (Chazal et al., 2011a) and (Caillerie et al., 2011), it is possible to extend our results to persistence diagram estimation for data corrupted by different kinds of noise using Wasserstein deconvolution methods. In another direction, an interesting representation of persistence diagrams as elements of a Hilbert space has recently been proposed in (Bubenik, 2012). Our results easily extend to this representation called *persistence landscapes*. Following this point of view, we also intend to adapt classical kernel-based methods with kernels carrying topological information.

References

Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., and Davis, J. SCAPE: shape completion and animation of people. In *Proceedings of SIGGRAPH*, pp. 408–416, 2005.

Balakrishnan, S., Rinaldo, A., Sheehy, D., Singh, A., and Wasserman, L. A. Minimax rates for homology inference. *Journal of Machine Learning Research - Proceedings Track*, 22:64–72, 2012.

Balakrishnan, S., Fasy, B.and Lecci, F., Rinaldo, A., Singh, A., and Wasserman, L. Statistical inference for persistent homology. *arXiv preprint arXiv:1303.7117*, 2013.

Bubenik, P. Statistical topology using persistence landscapes. *ArXiv e-prints*, July 2012.

Bubenik, P. and Kim, P. T. A statistical approach to persistent homology. *Homology, Homotopy and Applications*, 9(2):337–362, 2007.

Burago, D., Burago, Y., and Ivanov, S. *A course in metric geometry*, volume 33. Amer. Math. Soc., 2001.

Caillerie, C., Chazal, F., Dedecker, J., and Michel, B. Deconvolution for the Wasserstein metric and geometric inference. *Electronic Journal of Statistics*, 5:1394–1423, 2011.

Carlsson, G. Topology and data. *AMS Bulletin*, 46(2):255–308, 2009.

Chazal, F., Cohen-Steiner, D., Glisse, M., Guibas, L.J., and Oudot, S.Y. Proximity of persistence modules and their diagrams. In *SCG*, pp. 237–246, 2009a. ISBN 978-1-60558-501-7. doi: http://doi.acm.org/10.1145/1542362.1542407.

Chazal, F., Cohen-Steiner, D., Guibas, L. J., Memoli, F., and Oudot, S. Y. Gromov-Hausdorff stable signatures for shapes using persistence. *Computer Graphics Forum*, pp. 1393–1403, 2009b.

Chazal, F., Cohen-Steiner, D., and Mérigot, Q. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11(6):733–751, 2011a.

Chazal, F., Guibas, L. J., Oudot, S., and Skraba, P. Persistence-based clustering in Riemannian manifolds. In *Proc. of the 27th ACM Symp. on Computational Geometry*, 2011b.

Chazal, F., de Silva, V., Glisse, M., and Oudot, S. The structure and stability of persistence modules. *arXiv preprint arXiv:1207.3674*, 2012.

Chazal, F., de Silva, V., and Oudot, S. Persistence stability for geometric complexes. *Geometriae Dedicata (to appear - online first Dec. 2013)*, 2013a.

Chazal, Frédéric, Glisse, Marc, Labruère, Catherine, and Michel, Bertrand. Optimal rates of convergence for persistence diagrams in topological data analysis. *arXiv preprint arXiv:1305.6239*, 2013b.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. Stability of persistence diagrams. In *SCG*, pp. 263–271, 2005.

Cuevas, A. Set estimation: another bridge between statistics and geometry. *Bol. Estad. Investig. Oper.*, 25(2): 71–85, 2009. ISSN 1889-3805.

Cuevas, A. and Fraiman, R. A plug-in approach to support estimation. *Ann. Stat.*, 25(6):2300–2312, 1997.

Cuevas, A. and Rodríguez-Casal, A. On boundary estimation. *Adv. in Appl. Prob.*, 36(2):340–354, 2004.

De Silva, V. and Ghrist, R. Homological sensor networks. *Notices of the American Math Soc*, 54(1), 2007.

De Vito, E., Rosasco, L., and Toigo, A. Learning sets with separating kernels, 2012. URL http://arxiv.org/abs/1204.3573. submitted.

Edelsbrunner, H. The union of balls and its dual shape. *Discrete Comput. Geom.*, 13(1):415–440, 1995.

Edelsbrunner, H. and Harer, J. *Computational topology: an introduction*. American Math. Soc., 2010.

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

Genovese, C., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. Minimax manifold estimation. *Journal of Machine Learning Research*, 13:1263–1291, july 2012a.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. Manifold estimation and singular deconvolution under hausdorff loss. *Ann. Statist.*, 40:941–963, 2012b.

Hatcher, A. *Algebraic Topology*. Cambridge Univ. Press, 2001.

https://sites.google.com/site/nips2012topology/. Nips 2012 workshop on alg. topol. and machine learning.

Kasson, P. M., Zomorodian, A., Park, S., Singhal, N., Guibas, L. J., and Pande, V. S. Persistent voids: a new structural metric for membrane fusion. *Bioinformatics*, 23(14):1753–1759, 2007.

Korostelëv, A. P. and Tsybakov, A. B. *Minimax theory of image reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1993.

Mileyko, Y., Mukherjee, S., and Harer, John. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12), 2011.

Niyogi, P., Smale, S., and Weinberger, S. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39 (1-3):419–441, March 2008.

Niyogi, P., Smale, S., and Weinberger, S. A topological view of unsupervised learning from noisy data. *SIAM Journal on Computing*, 40(3):646–663, 2011.

Rodríguez-Casal, A. Set estimation under convexity type assumptions. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 43(6):763 – 774, 2007.

Singh, A., Scott, C., and Nowak, R. Adaptive Hausdorff estimation of density level sets. *Ann. Stat.*, 37(5B):2760–2782, 2009.

Singh, G., Memoli, F., Ishkhanov, T., Sapiro, G., Carlsson, G., and Ringach, D. L. Topological analysis of population activity in visual cortex. *Journal of vision*, 8(8), 2008.

Tsybakov, A. B. On nonparametric estimation of density level sets. *Ann. Statist.*, 25(3):948–969, 1997.

Wang, J. *Geometric structure of high-dimensional data and dimensionality reduction*. Springer, 2012.

Zomorodian, A. and Carlsson, G. Computing persistent homology. *Disc. Comp. Geom.*, 33(2):249–274, 2005.