Statistical learning on measures: an application to persistence diagrams

Olympio Hacquard

Statistical learning on measures: an application to persistence diagrams

Olympio Hacquard OLYMPIO.HACQUARD@UNIVERSITE-PARIS-SACLAY.FR
Laboratoire de Mathématiques d'Orsay
307 rue Michel Magat,
Université Paris-Saclay, Inria
91400 Orsay France

Editor:

Abstract

We consider a binary supervised learning classification problem where instead of having data in a finite-dimensional Euclidean space, we observe measures on a compact space $\mathcal{X}$. Formally, we observe data $D_N = (\mu_1, Y_1), \ldots, (\mu_N, Y_N)$ where μ_i is a measure on $\mathcal{X}$ and Y_i is a label in $\{0, 1\}$. Given a set $\mathcal{F}$ of base-classifiers on $\mathcal{X}$, we build corresponding classifiers in the space of measures. We provide upper and lower bounds on the Rademacher complexity of this new class of classifiers that can be expressed simply in terms of corresponding quantities for the class $\mathcal{F}$. If the measures μ_i are uniform over a finite set, this classification task boils down to a multi-instance learning problem. However, our approach allows more flexibility and diversity in the input data we can deal with. While such a framework has many possible applications, this work strongly emphasizes on classifying data via topological descriptors called persistence diagrams. These objects are discrete measures on $\mathbb{R}^2$, where the coordinates of each point correspond to the range of scales at which a topological feature exists. We will present several classifiers on measures and show how they can heuristically and theoretically enable a good classification performance in various settings in the case of persistence diagrams.

Keywords: Statistical learning, Measures, Topological data analysis, Persistence diagrams, Classification.

1. Introduction

We consider the problem of classifying measures over some metric space. This problem appears as a generalization of standard supervised classification where the data are no longer vectors from a Euclidean space but point clouds or even continuous measures. There are several lines of work looking at this problem from various perspectives. If the measure is a finite sum of Dirac masses, this problem boils down to multi-instance learning (MIL), where the data are bags of points. This terminology originates in the works from Dietterich et al. (1997) in the context of drug design. A typical strategy in MIL is to consider a standard classifier over the points from the bag and aggregate the individual labels to classify the entire bag. We refer to the survey from Amores (2013) for a comprehensive review of the methods used in multi-instance classification. Closer to our work is the paper by Sabato and Tishby (2012), which studies the properties of MIL from a statistical learning perspective. More general are the works on distribution regression, for instance, Muandet

et al. (2012) for classification and Póczos et al. (2013) for regression. The theory for simple kernel estimators has been developed in Szabó et al. (2016). Another recent perspective regarding distribution learning follows the works by Moosmüller and Cloninger (2020) and Khurana et al. (2023), where the authors consider that each class consists of perturbations of a "mother distribution" and tackle this problem using tools from optimal transport. To conclude our overview of measure-learning methods, we can cite the work from Chazal et al. (2021), where the authors vectorize the measures to cluster them or perform a supervised learning task. The setting we consider is very general in the type of measures we handle, and vectorization-free. We consider simple classifiers based on integrals over the sample measure, and we look at the theoretical performance of such classifiers by relating complexity measures such as Rademacher complexity and covering numbers to their counterparts in the base space. This follows a similar approach as Sabato and Tishby (2012), while we allow for more general inputs. We can therefore derive standard prediction bounds, see Mohri et al. (2018) for an introduction to these concepts. We provide classification algorithms which fit under this framework and that discriminates according to the fraction of the mass each measure puts in a well-chosen area.

The theory developed here has many cases of applications, namely, whenever input data are point clouds. We can, for instance, cite lidar reconstruction (De Deuge et al. (2013)), flow cytometry (Aghaeepour et al. (2013)), time series (possibly with an embedding mapping them in some Euclidean space), and text classification using a word embedding method such as WORD2VEC, see Mikolov et al. (2013). Extending the results from MIL, the measures can be weighted depending on the application. We also encompass the case of continuous measures, for example, functional or image classification.

The main application that motivates the present work is the classification of persistence diagrams. We refer to Edelsbrunner and Harer (2022) for an overview of this object's construction and its principal properties. Persistence diagrams are stable topological descriptors of the filtration of a simplicial complex. Mathematically, they are discrete measures on $\mathbb{R}^2$ where both coordinates of each point indicate times at which topology changes occur in the filtration. We can use persistence diagrams to perform various data analysis tasks, and we focus here on supervised classification to discriminate data based on some topological information. Some methods such as landscapes (Bubenik et al. (2015), persistence images (Adams et al. (2017)), or ATOL (Chazal et al. (2021)) immediately get rid of the measure representation and transform the data in a vector. It then becomes possible to plug a standard classifier into these vector representations, we refer to Obayashi et al. (2018) for classification using linear classifiers. Some papers use kernel methods, such as Carriere et al. (2017) or Le and Yamada (2018), while some other works make use of neural networks, such as Carrière et al. (2020), and more recently Reinauer et al. (2021). We refer to the survey Hensel et al. (2021) for an overview of topological machine learning methods.

In addition to offering a good trade-off between decent predictive performance (comparable to standard persistence diagrams vectorizations and kernel methods) and simplicity, the algorithm developed here offers explainability guarantees. Indeed, showing that two classes differ on some zones of the persistence diagrams can directly be translated in terms of the range of scales at which relevant topological features exist. The experimental results back up the ideas developed by Bubenik et al. (2020) by swiping away a typical paradigm in topological data analysis (TDA), which states that features with a long lifetime are the

only ones relevant to describing a shape. Indeed, we demonstrate that the "shape" of the topological noise contains information related to the sampling. This idea is enforced by theoretical guarantees on limiting persistence diagrams as the number of sample points tends to infinity, where we generalize recent results from Owada (2021).

The present work decomposes as follows: in Section 2 we formalize the problem of learning on a set of measures, give general theoretical guarantees for this problem and propose two simple supervised algorithms. In Section 3, we present persistence diagrams that constitute the primary motivation of the present work. We give guarantees on the reconstruction of the proposed algorithm in this specific case, showing that features at every scale can and should be used for classification. Section 4 contains all the experimental results and the comparison with standard methods, both in TDA and for other applications, showing the versatility of our approach, which we believe is its principal strength, along with its simplicity and explainability. We have made the code available here [1]. Finally, Section 5 is devoted to the proofs of all the theoretical results contained throughout the paper.

2. Statistical learning on measures

2.1 Model

Let $\mathcal{X}$ be a compact metric space and denote by $\mathcal{M}(\mathcal{X})$ the set of measures of finite mass over $\mathcal{X}$. Assume that both $\mathcal{X}$ and $\mathcal{M}(\mathcal{X})$ are endowed with σ-algebras. This rather abstract assumption is such that we can consider random variables on $\mathcal{X}$ and $\mathcal{M}(\mathcal{X})$. The model is the following: we observe a sample $D_N = (\mu_i, Y_i)_{i=1}^N$, where $\mu_i \in \mathcal{M}(\mathcal{X})$ and Y_i is a label in $\mathcal{Y} \subset \mathbb{R}$. Although the algorithmic and experimental study is mainly motivated by the case of classification $\mathcal{Y} = \{0, 1\}$, some of the theory developed also encompasses the case of regression where $\mathcal{Y} = [0, 1]$. We aim at building a decision rule $g : \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ that predicts the label Y' of a new measure μ'. These decision rules are typically built on classes of functions defined on $\mathcal{X}$ itself. There are many practical examples that fall under this framework of learning on a space of measures: functional regression (Ferraty and Vieu (2006)) and image classification are standard examples that have given birth to a very wide variety of problems. Classifying bags of points has been studied under the MIL terminology, we refer to Amores (2013) for a complete survey, and cover many useful applications, from which we can cite image classification based on a finite number of descriptors as done in Wu et al. (2015), flow cytometry (see Section 4), or text classification where each word is represented by a point in a high-dimensional space. Closer to us is the work by Chazal et al. (2021) where they represent the measures in a Euclidean space and use these vectorizations to cluster the data. Even though the applications are very similar, we believe that our work is quite different in essence since our algorithms are formulated in a supervised setting, and we do not represent the measures in a Euclidean space, preferring to develop a theory directly for an input space of measures, as we will see in the following section.

[1] `https://github.com/OlympioH/BBA_measures_classification`

2.2 Theoretical complexity bounds

In this section, we adapt standard results in statistical learning theory by relating quantities such as Rademacher complexity and covering numbers for functional classes over $\mathcal{X}$ to their counterparts in the space $\mathcal{M}(\mathcal{X})$. In what follows, $\mathcal{R}_N(\cdot)$ (resp. $G_N(\cdot)$) denotes the empirical Rademacher (resp. Gaussian) complexity of a function class on a sample $(Z_1, \dots, Z_N)$ which we recall is defined by

$$\mathcal{R}_N(\mathcal{F}) = \frac{1}{N}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{N}\sigma_i f(Z_i)\right|\right],$$

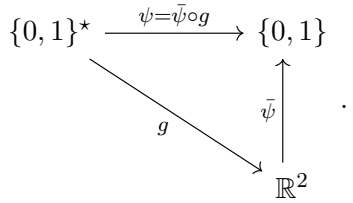
where $(\sigma_1, \dots, \sigma_N)$ are independent Rademacher random variables. The Gaussian complexity obeys the same definition where the σ_i are independent standard normal variables. The Rademacher complexity is a usual quantity in statistical learning that measures the richness of a set of functions. Loosely speaking, it quantifies how much the class $\mathcal{F}$ correlates with a vector of noise $(\sigma_1, \dots, \sigma_N)$. This quantity naturally appears when controlling the performance of a family of classifiers; a large Rademacher complexity being detrimental to a good generalization. We refer to Chapter 26 from Shalev-Shwartz and Ben-David (2014) for more details. It is common to upper bound it by computing the covering number of the function class. We denote by $\mathcal{N}(\mathcal{F}, d, \varepsilon)$ (resp. $\mathcal{M}(\mathcal{F}, d, \varepsilon)$) the ε-covering (resp. packing) number of the set $\mathcal{F}$ endowed with metric d. Finally, we denote by $vc(\mathcal{F})$ the Vapnik-Chervonenkis complexity of a set of functions (or its pseudo-dimension in the case of real hypotheses classes) and by $vc(\mathcal{F}, \gamma)$ its γ-fat shattering dimension. We refer to Chapter 6 of Shalev-Shwartz and Ben-David (2014) for the definition of these concepts that measure the capacity of a function class, and are also used to upper bound the validation error of a classification model. We break down our analysis in two cases: the first one assumes that we have discrete finite measures and that we apply the 0-1 loss while the second assumes generic measures inputs and requires the loss function to be Lipschitz.

2.2.1 Discrete measures, 0-1 loss

We denote by $\mathcal{M}_m(\mathcal{X})$ the set of measures that write as a finite sum of at most m Dirac masses on $\mathcal{X}$, i.e. $\mu_i = \sum_{j=1}^{n_i}\delta_{x_j^i}$ with $n_i \leq m$ for all i. We consider a family $\mathcal{F}$ of classifiers from $\mathcal{X}$ to $\{0,1\}$. For a given $f \in \mathcal{F}$, we have a set of predictions for each individual point:

$$f(\mu_i) = [f(x_1^i), f(x_2^i), \dots, f(x_{n_i}^i)] \in \{0,1\}^{n_i}.$$

Denoting by $\{0,1\}^\star$ the set of finite sequences of 0's and 1's, we finally apply some function $\psi : \{0,1\}^\star \to \{0,1\}$ called a bag-function or an aggregation function in order to output a prediction label for each measure. Described as such, this scenario is formulated exactly as a Multi-Instance Learning (MIL) problem, and theoretical guarantees in this case have been established by Sabato and Tishby (2012). In Proposition 1, we extend their results, in particular their Theorem 6 to the case where ψ is no longer a fixed-function but is itself learned from a VC-class $\mathcal{G}$. Let the bag-function ψ be permutation invariant, i.e. $\psi(y_1, \dots, y_n) = \psi(y_{\sigma(1)}, \dots, y_{\sigma(n)})$ for every $y_i \in \{0,1\}, n \in \mathbb{N}$, and $\sigma \in \mathfrak{S}_n$. There exist two functions g and $\bar{\psi}$ such that ψ decomposes as follows:



The function g is defined as $g(y_1, \ldots, y_n) = (\sum_{i=1}^n y_i/n, n)$, i.e. it maps a sequence of 0's and 1's to the proportion of 1's and the total number of elements in the sequence. We denote by $\mathcal{H}$ the set of binary classifiers from $\mathcal{M}_m(\mathcal{X})$ defined as $h : \mu = \sum_{i=1}^n \delta_{x_i} \mapsto \psi(f(x_1), \ldots, f(x_n))$, where $\psi \in \mathcal{G}$ and $f \in \mathcal{F}$.

Proposition 1 *Assume all the input measures belong to $\mathcal{M}_m(\mathcal{X})$. Assume ψ is taken from a class $\mathcal{G}$ of permutation invariant functions and that the corresponding $\bar{\psi}$ is taken from a class $\bar{\mathcal{G}}$ of VC-dimension d'. We further assume that the class $\mathcal{F}$ has a finite VC-dimension d. Then, $\mathcal{H}$ is a VC-class of dimension d_2 verifying:*

$$d_2 \leq \max(16, (d + d') \log_2(2em)).$$

We defer the proof to Section 5.1. This bound on the VC dimension of the composition of a hypothesis class $\mathcal{F}$ with a class of bag-functions can be used to upper-bound the classification accuracy of predictors over the set of measures. We now propose to extend these results to the case of general measures with finite mass and therefore extend the MIL framework.

2.2.2 Generic measures, Lipschitz loss

In this subsection, we build classifiers of the form $\mathrm{sgn}(g(\mu))$ for g in some function class $\mathcal{G}$. Consider a κ-Lipschitz loss function $\mathcal{L}$. By the contraction principle, $\mathcal{R}_N(\mathcal{L} \circ \mathcal{G}) \leq \kappa \mathcal{R}_N(\mathcal{G})$. This subsection is therefore devoted to the control of the Rademacher complexity of the class of real-valued predictors. We first extend Lemma 12 from Sabato and Tishby (2012) to our setting. In what follows, we consider a class of functions $\mathcal{F}$ from $\mathcal{X}$ to $[0, 1]$, and the associated class of functions $\tilde{\mathcal{F}}$ defined on $\mathcal{M}(\mathcal{X})$ by

$$\tilde{f}[\mu] = \mathbb{E}_{X \sim \mu}[f(X)] = \int_{\mathcal{X}} f(x) \mathrm{d}\mu(x) \text{ for } f \in \mathcal{F}.$$

The following lemma gives a relationship between the covering numbers of $\mathcal{F}$ and $\tilde{\mathcal{F}}$. We denote by L_p^N the in-sample norm, defined as such for two functions f_1 and f_2 in $\mathcal{F}$ and a sample of N measures $(\mu_1, \ldots, \mu_N)$:

$$\|\tilde{f}_1 - \tilde{f}_2\|_{L_p^N} = \left(\frac{1}{N} \sum_{i=1}^N (\tilde{f}_1[\mu_i] - \tilde{f}_2[\mu_i])^p \right)^{1/p}.$$

Given a sample $(\mu_1, \ldots, \mu_N) \in \mathcal{M}(\mathcal{X})^N$, we denote by $\bar{M}_p = \left(\frac{1}{N} \sum_{i=1}^N M_i^p \right)^{1/p}$ where $M_i = \mu_i(\mathcal{X})$ is the total mass of the measure μ_i.

Lemma 2 *Let $(\mu_1, \dots, \mu_N) \in \mathcal{M}(\mathcal{X})^N$ and let $p \in [1, +\infty[$. There exists a probability measure $\bar{\mu}$ such that*

$$\mathcal{N}(\tilde{\mathcal{F}}, L_p^N, \varepsilon) \leq \mathcal{N}\left(\mathcal{F}, L_p(\bar{\mu}), \frac{\varepsilon}{\bar{M}_p}\right).$$

We defer the proof to Section 5. This can be used to upper-bound the Rademacher complexity of the function class $\tilde{\mathcal{F}}$ as shown in the following theorem.

Theorem 3 *There exists an absolute constant K such that*

$$\mathcal{R}_N(\tilde{\mathcal{F}}) \leq \frac{K\bar{M}_2\sqrt{vc(\mathcal{F})}}{\sqrt{N}}.$$

For $M > 0$, we define the set $\mathcal{C}_M^N = \{(\mu_1, \dots, \mu_N) \in (\mathcal{M}(\mathcal{X}))^N | \frac{1}{N}\sum_{i=1}^N \mu_i^2(\mathcal{X}) = M^2\}$. In addition to the previous theorem, we provide a lower bound of the same order for the Rademacher complexity.

Theorem 4 *There exists an absolute constant K' such that*

$$\frac{K'\bar{M}_2}{\sqrt{N}\ln(N)}\sqrt{vc(\mathcal{F})} \leq \sup_{(\mu_1,\dots,\mu_N)\in\mathcal{C}_{\bar{M}_2}^N} \mathcal{R}_N\left(\tilde{\mathcal{F}}|\mu_1, \dots, \mu_N\right).$$

The bounds from Theorems 3 and 4 match and are both of order $1/\sqrt{N}$, up to log-factors. They also both depend on the VC-dimension of the base-class $\mathcal{F}$ and no longer of $\tilde{\mathcal{F}}$, making it much easier to compute, as we can see in the example below.

Example 1 *Assume $\mathcal{X}$ is a bounded subspace of $\mathbb{R}^d$ endowed with a Euclidean metric and let $\mathcal{F} = \{\mathbb{1}_{\mathcal{B}(x,r)} | x \in \mathcal{X}, r > 0\}$. It is a standard fact (see Mohri et al. (2018) for instance) that the VC-dimension of Euclidean balls is $d + 1$. We therefore have by Theorem 3 that there exist constants K and K' such that:*

$$\frac{K'\bar{M}_2\sqrt{d+1}}{\sqrt{N}\ln(N)} \leq \sup_{(\mu_1,\dots,\mu_N)\in\mathcal{C}_{\bar{M}_2}^N} \mathcal{R}_N\left(\tilde{\mathcal{F}}|\mu_1, \dots, \mu_N\right) \leq \frac{K\bar{M}_2\sqrt{d+1}}{\sqrt{N}}.$$

In practice, the class $\tilde{\mathcal{F}}$ is used to construct a binary classifier through composition with an aggregation function ψ, whose sign gives a prediction in $\{0, 1\}$. If the function ψ is fixed as it is the case in Sabato and Tishby (2012) and is further assumed to be L-Lipschitz, the Rademacher complexity of the final set of classifiers is simply multiplied by L. We want to generalize this to the case where the function ψ is also learned. Assume $\psi : \mathbb{R} \to \mathbb{R}$ is taken from a class of functions $\mathcal{G}$. Denote by $\mathcal{H}$ the class of functions $h : \mu \mapsto \psi(\int_{\mathcal{X}} f(x)\mathrm{d}\mu(x))$ where $f \in \mathcal{F}, \psi \in \mathcal{G}$. The following proposition gives a bound on the Gaussian complexity of the function class $\mathcal{H}$.

Proposition 5 *Assume that the class $\mathcal{G}$ consists of L-Lipschitz functions. Assume the function $x \mapsto 0$ belongs to $\mathcal{F}$. Then there exist constants C_1 and C_2 such that for any sample of measures $\bar{\mu} = (\mu_1, \dots, \mu_N)$,*

$$G_N(\mathcal{H}) \leq \frac{C_1 \bar{M}_2 L \sqrt{vc(\mathcal{F})}\sqrt{\log(N)}}{\sqrt{N}} + \frac{C_2 L \bar{M}_2 \mathbf{R}(\mathcal{G})}{\sqrt{N}} + \frac{L}{\sqrt{N}} \sup_{\psi \in \mathcal{G}} |\psi(0)|,$$

where

$$\mathbf{R}(\mathcal{G}) = \sup_{\mathbf{x}, \mathbf{x}' \in \mathbb{R}, \mathbf{x} \neq \mathbf{x}'} \mathbb{E}_\gamma \sup_{\psi \in \mathcal{H}} \frac{(\psi(\mathbf{x}) - \psi(\mathbf{x}'))\gamma}{|\mathbf{x} - \mathbf{x}'|},$$

and where $\gamma \sim \mathcal{N}(0, 1)$.

We refer to Section 5.5 for the proof. This proposition shows that up to logarithmic factors, the Gaussian complexity of the family of classifiers decreases at an overall rate of $1/\sqrt{N}$. The quantity $\mathbf{R}(\mathcal{G})$ appears as a supremum of Gaussian averages. We refer to Theorem 5 of Maurer (2016) for a few properties of this quantity. Most notably, if the class $\mathcal{G}$ is finite, and consists of L-Lipschitz functions, $\mathbf{R}(\mathcal{G}) \leq L\sqrt{2 \ln |\mathcal{G}|}$. In addition, in some simple cases, it is possible to provide a better estimate of $\mathbf{R}(\mathcal{G})$, even when $\mathcal{G}$ is infinite, as we can see in the following example:

Example 2 *In practice, we often choose ψ of the form $\psi : x \mapsto x - s$ where $s \in [-S, S]$ is learned, as we will see in Section 2.3. In this case, we directly have that $\mathbf{R}(\mathcal{G}) = \mathbb{E}[|\gamma|] = 1$. Therefore, keeping the same notation as above, there exist universal constants C_1 and C_2 such that*

$$G_n(\mathcal{H}) \leq \frac{1}{\sqrt{N}} \left[C_1 \bar{M}_2 \sqrt{vc(\mathcal{F})}\sqrt{\log(N)} + S + C_2 \bar{M}_2 \right].$$

2.3 Algorithms, application to rectangle-based classification

Let us consider a class $\mathcal{A}$ of Borel sets of $\mathcal{X}$. For instance, $\mathcal{A}$ can be thought of as the set of balls or axis-aligned hyperrectangles for a given metric. We then consider the class of corresponding indicator functions $\mathcal{F} = \{\mathbb{1}_A, A \in \mathcal{A}\}$. The data are therefore classified given some threshold $s \in \mathbb{R}$ and a sign $\varepsilon \in \{-1, +1\}$, by the decision rule $\mu \mapsto \mathbb{1}\{\varepsilon \int_A \mathrm{d}\mu - s \geq 0\}$.

If $\mathcal{A}$ is a set of balls, the optimization problem boils down to finding the best center in $\mathcal{X}$ and the best radius in $\mathbb{R}_+$. We present two algorithms and associate each of them with the theory developed in the previous subsection.

Algorithm 1: exhaustive search

The first method consists in performing an exhaustive search in a grid of parameters for a threshold $s \geq 0$, and take the ones that minimize the empirical classification error:

$$\mathcal{L}_+ = \sum_{i=1}^{N} \mathbb{1}\left\{ \int_A \mathrm{d}\mu_i - s > 0 \right\} \mathbb{1}\{Y_i = 0\} + \mathbb{1}\left\{ \int_A \mathrm{d}\mu_i - s \leq 0 \right\} \mathbb{1}\{Y_i = 1\}.$$

We then minimize the empirical classification error for reversed labels:

$$\mathcal{L}_- = \sum_{i=1}^{N} \mathbb{1}\left\{\int_A \mathrm{d}\mu_i - s \leq 0\right\} \mathbb{1}\{Y_i = 0\} + \mathbb{1}\left\{\int_A \mathrm{d}\mu_i - s > 0\right\} \mathbb{1}\{Y_i = 1\}.$$

If $\mathcal{L}_+ \leq \mathcal{L}_-$ we set $\varepsilon = 1$ and $\varepsilon = -1$ otherwise, along with the corresponding set of parameters.

If all the μ_i's write as a finite sum of Dirac measures, this method is very similar to MIL, since each of the N_i points in the bag μ_i will be assigned a label whether it belongs to the set A or not. The slight difference is that we consider multiple aggregation functions of the form

$$\begin{aligned} \psi_s \colon \{0,1\}^{N_i} &\to \{0,1\} \\ x &\mapsto \mathbb{1}\left\{\sum_{j=1}^{N_i} x_j \geq s\right\}. \end{aligned}$$

Here, we allow the threshold s to be learned, which slightly extends the theory developed in Chapter 3 of Sabato and Tishby (2012) about binary MIL where the aggregation function must be fixed. We therefore fit exactly within the framework of Proposition 1 provided that the set of raw classifiers $\mathcal{F} = \{\mathbb{1}_A | A \in \mathcal{A}\}$ is a VC-class, which is for instance the case if $\mathcal{A}$ is a set of Euclidean balls or axis-aligned hyperrectangles. Note that this algorithm allows for any sample of measures with finite mass as input.

Algorithm 2: smoothed version

Performing an exhaustive search has a computational cost that grows exponentially with the dimension of the space in which the data lie. We propose to optimize a smoothed version of the empirical error. In the case of balls, for a center $C \in \mathcal{X}$, a radius $r > 0$, a threshold s and a scale σ, we consider the predictor given by the sign of $f_{C,r,s,\sigma}$, defined as

$$f_{C,r,s,\sigma}(\mu) = \int_{\mathcal{X}} \exp\left(-\frac{d(\mathcal{B}(C,r),x)}{\sigma}\right) \mathrm{d}\mu(x) - s.$$

We minimize the cross-entropy loss between a smooth version of this predictor and the target vector, for a sample $D_N = (\mu_i, Y_i)_{i=1}^N$:

$$\mathcal{L}_{D_N}(C,r,s,\sigma) = -\sum_{k=1}^{N} Y_k \log(S(f_{C,r,s,\sigma}(\mu_k)) + (1-Y_k)\log(1 - S(f_{C,r,s,\sigma}(\mu_k)),$$

where S is the sigmoid function: $x \mapsto \frac{1}{1+e^{-x}}$. This optimization must be performed for switched labels as well.

In practice, we perform a stochastic gradient descent of this loss function. Since this objective typically has many critical points, we perform multiple runs with different initialization parameters.

The predictor $S \circ f_{C,r,s,\sigma}$ is a smooth predictor that has output in $\mathcal{Y} = [0,1]$. This algorithm can also be interpreted using the MIL lens if the μ_i's are discrete sums of Dirac measures. Indeed, the class of functions we consider is

$$\mathcal{F} = \left\{ x \mapsto \exp\left(-\frac{d(\mathcal{B}(C,r),x)}{\sigma}\right) | C \in \mathcal{X}, (r,\sigma) \in (\mathbb{R}_+)^2 \right\},$$

so that each point in the bag μ_i is mapped to a real number which corresponds to the framework of Section 6.2 of Sabato and Tishby (2012). The class $\mathcal{F}$ is a smoothed version of balls' indicators and has the same VC dimension: $vc(\mathcal{F}) = d+1$. We can therefore write the corresponding generalization bound, using that the the cross-entropy loss is 1-Lipschitz. According to Theorem 26.5 of Shalev-Shwartz and Ben-David (2014), we have that with probability at least $1-\delta$, for all $\theta = (C, r, \sigma, s) \in \mathcal{X} \times \mathbb{R}_+^2 \times [-S, S]$,

$$\mathbb{E}_S[\mathcal{L}_S(\theta)] - \mathcal{L}_{D_N}(\theta) \leq \sqrt{\frac{\pi}{2N}} \left[C_1 \sqrt{(d+1)\log(N)} \bar{M}_2 + S + C_2 \bar{M}_2 + C_3 \sqrt{\log(4/\delta)} \right],$$

with universal constants C_1, C_2 and C_3.

Aggregation phase

These two methods select a single Borel set to discriminate between the two classes. This approach suffices when the two classes always differ in the same zone of $\mathcal{X}$. However, this only corresponds to the simpler applications, and we want to find several disjoint Borel sets along with corresponding thresholds of activation and aggregate them. We therefore propose a boosting approach to combine "weak-classifiers" to build a strong one. We have implemented the ADABOOST method developed for the first time by Freund et al. (1996). It follows an iterative procedure, giving more and more weight to misclassified data. In addition to greatly improving the predictive performance as opposed to selecting a single convex set, performing a boosting is of qualitative interest since it shows which zones of the measures are relevant for classification. This feature is of particular interest in applications where these areas convey a qualitative information, such as persistence diagrams or flow cytometry.

3. A leading case study: classifying persistence diagrams

The primary example of measures that motivates the present work are persistence diagrams and their smoothed and weighted variants.

3.1 An introduction to persistence diagrams

Persistence diagrams are measures on $\mathbb{R}^2$ that summarize the topological properties of input data and constitute one of the main objects in Topological Data Analysis (TDA). We refer to the textbooks by Edelsbrunner and Harer (2022) and Boissonnat et al. (2018) for complete and exhaustive treatment, and recall the principal notions. Given a nested sequence $(X_t)_{t\in\mathbb{R}}$ of topological spaces built on top of the data, we keep track of the evolution of the topology as t grows from $-\infty$ to $+\infty$. More precisely, we look at k-dimensional homological features (connected components in dimension 0, cycles in dimension 1, voids in dimension 2, and so

on). If a k-dimensional feature is born at time b and dies at time d, we add the point (b, d) in the k-dimensional persistence diagram ξ_k. The latter therefore appears as a multi-set of points supported in the half-plane H defined by $H = \{(x, y) \in \mathbb{R}^2 |\ x \leq y \leq +\infty\}$. We can equivalently look at persistence diagrams as discrete measures on H: $\xi_k = \sum_{(b,d)} \delta_{(b,d)}$. We usually add the diagonal $\Delta = (x, x) \subset H$ to the persistence diagrams in order to compare diagrams with a different number of points. To do so, we use a transport distance called the bottleneck distance:

Definition 6 *The bottleneck distance d_B between two persistence diagrams D_1 and D_2 is defined by:*

$$d_B(D_1, D_2) = \inf_{\eta: D_1 \to D_2} \sup_{x \in D_1} \|x - \eta(x)\|_\infty,$$

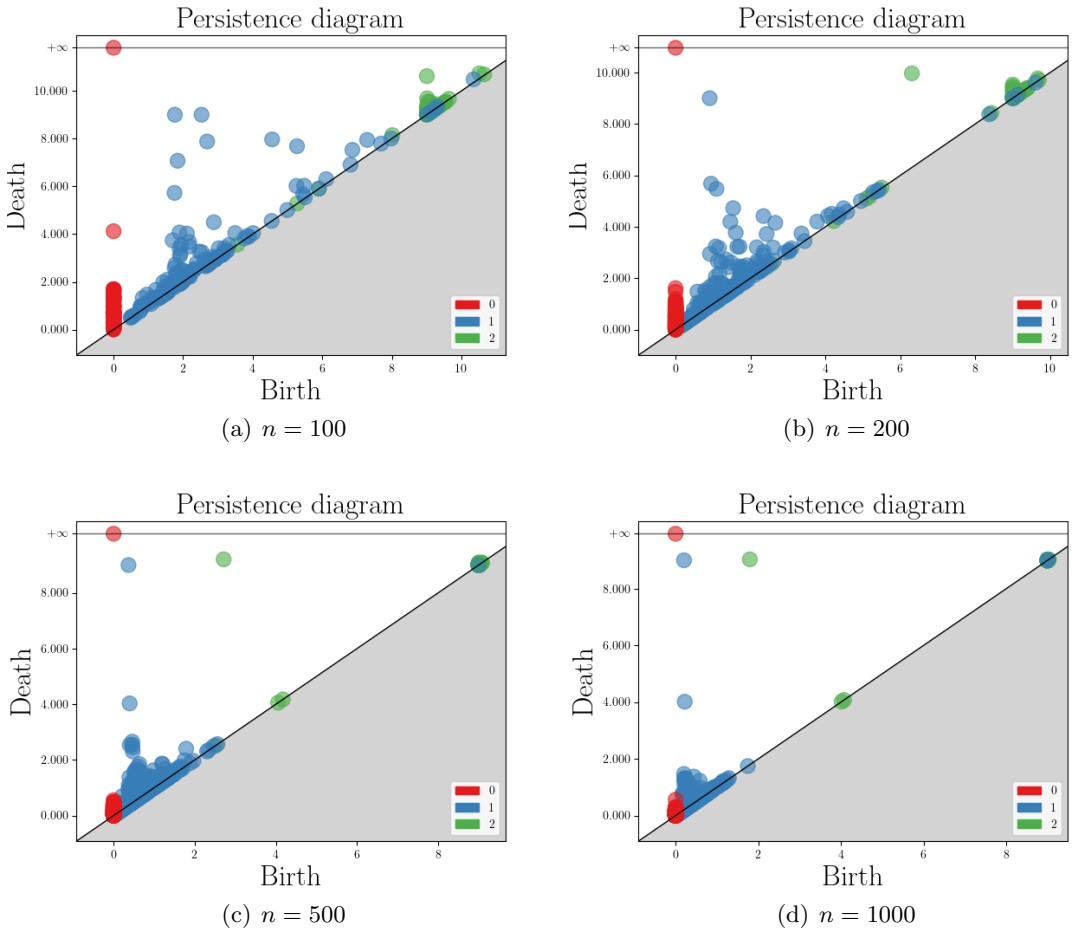
where the infimum is taken over all bijections η from D_1 to D_2.

From now on, we will focus on point cloud data unless specified otherwise, and all the diagrams are built using the Cech filtration:

Definition 7 *For a point cloud $\mathbb{X} = (X_i)_{i=1}^n$, The Cech complex $\check{C}(\mathbb{X}, r)$ at scale r is the simplicial complex where for every simplex $\sigma = [X_{i_0}, \ldots, X_{i_k}]$, we have $\sigma \in \check{C}(\mathbb{X}, r)$ if $\cap_{i=i_0}^{i_k} \mathcal{B}(X_i, r) \neq \emptyset$.*

We then let r go from 0 to ∞ and track the evolution of the homology of the union of balls. For a given point (b, d) in the persistence diagram, its distance to the diagonal $d-b$ is called its persistence or lifetime and measures for how long this feature persists in the filtration.

We illustrate this construction in Figure 1 where we sample n points uniformly on a torus, according to an algorithm provided by Diaconis et al. (2013). When the number of points is very low ($n = 100$), the true homology of the manifold (one feature of dimension 2 and two features of dimension 1) does not show in the diagrams and we only observe topological components due to the sampling. For $n = 500$, we can read the homology of the torus in the persistence diagram along with many points close to the diagonal. As n grows, this "topological noise" concentrates around the origin and the true homological features become well separated from the noise. If we sample a point cloud from a manifold, large-persistence features correspond to proper homological features of the manifold, see Theorem 9. Following this approach, works such as Adams et al. (2017) on persistence images suggest weighting the persistence diagram using an increasing function of the persistence. In addition, they propose to convolve the discrete measure with a Gaussian function. This falls under the framework of the previous section, and it becomes relevant to consider diagrams as generic measures. However, this signal-noise dichotomy is very restrictive, and there is some evidence that points lying close to the diagonal also carry relevant information such as curvature as demonstrated in Bubenik et al. (2020) or dimension. We give further evidence of that in the following section, where we show that asymptotically, we can extract information on the sampling density around the origin of the limiting persistence diagram. We also provide numerical illustrations and quantitative evidence that low-persistence features are relevant for classification purposes.



(a) $n = 100$ (b) $n = 200$

(c) $n = 500$ (d) $n = 1000$

Figure 1: 0, 1 and 2-persistence diagrams for n points uniformly sampled on a torus.

3.2 Structural properties of persistence diagrams

Throughout this section and the following, we consider classifiers constructed by finding the best axis-aligned rectangle. The easiest information to capture on the persistence diagram of a Cech complex is the global one corresponding to the homology of the underlying manifold. If we consider samplings on metric spaces having different persistence diagrams for a given filtration, the following theorem yields the existence of a rectangle classifier that discriminates between the two supporting spaces with high probability. Before stating the theorem, we recall the definition of an (a, b)-standard measure:

Definition 8 *Let $\mathbb{X}$ be a compact metric space and let $a, b > 0$. We say that a probability measure μ on $\mathbb{X}$ satisfies the (a, b)-standard assumption if*

$$\forall x \in \mathbb{X}, \forall r > 0, \mu(\mathcal{B}(x, r)) \geq \min(1, ar^b).$$

Theorem 9 *Let $\mathbb{M}_1$ and $\mathbb{M}_2$ be two compact metric spaces. Assume that we observe an i.i.d. sample $\hat{X}_n = (X_i)_{i=1}^n$ drawn from an (a_1, b_1)-standard measure on $\mathbb{M}_1$ or an (a_2, b_2)-*

standard measure on $\mathbb{M}_2$. *Denote by* $K_i = \min_{(p,q)\in dgm(M_i)} \|(p,q)\|_\infty$ *for* $i = 1, 2$. *Assume that there exists* $K_3 > 0$ *such that*

$$d_B(dgm(\check{C}(\mathbb{M}_1)), dgm(\check{C}(\mathbb{M}_2))) \geq K_3.$$

Denote by $K = \min(K_1, K_2, K_3), a = \min(a_1, a_2)$ *and* $b = \min(b_1, b_2)$. *For all* $\delta > 0$, *if the number* n *of sample points verifies*

$$n \geq \frac{2^b}{aK^b} \log\left(\frac{4^b}{aK^b\delta}\right),$$

there exists a collection of rectangles for which the classification error is smaller than δ.

We defer the proof to Section 5.6. We refer to Chazal et al. (2014a) for the construction of simplicial complexes of possibly infinite metric spaces, which ensures that $dgm(\check{\mathrm{C}}(\mathbb{M}_1))$ and $dgm(\check{\mathrm{C}}(\mathbb{M}_2))$ are well defined.

In practice, a lot of information is contained in the points lying close to the diagonal and classifying persistence diagrams enables to deal with a far broader class of problems than simply classifying between manifolds with different homology groups, as we will see in the following sections. In order to investigate the impact of the sampling itself on persistence diagrams, we have derived the Theorem 10, where we study the asymptotic behavior of the persistent diagram of the Cech complex built on a point cloud sampled from a manifold. Before stating the theorem, we give a quick review of the existing literature on limiting theorems for persistent quantities for a random point cloud $(X_i)_{i=1}^n$ on $\mathbb{R}^d$. In what follows, we denote by β_k the k-th Betti number of a simplicial complex and N_k the number of critical simplices of dimension k. In terms of persistence diagrams, $\beta_k(\check{\mathrm{C}}(X, r))$ is the number of points in the upper left quadrant that intersect the diagonal at (r, r). It is well known that β_k and N_k are closely related by Morse theory: a critical simplex of dimension k can either increase β_k by 1 or decrease β_{k-1} by 1. We further denote by $\beta_k(r, s)$ the persistent Betti number between times r and s, i.e. the number of k-dimensional features that are born before r and die after s. It corresponds to the number of points in the upper-quadrant with an angle at the point (r, s) in the persistence diagram. Since we consider i.i.d. samples, we adopt the approach from Bobrowski and Kahle (2018) which gives a detailed survey on random geometric complexes that we enrich with a few recent results. Assume that we re-scale the data by a sequence r_n such that $r_n \to 0$. The speed at which r_n tends to 0 as $n \to \infty$ is paramount and we introduce the scale factor $\Lambda = nr_n^d$. There are three possible different behaviors according to the limit of Λ. We refer to Bobrowski and Adler (2011) for the results regarding the number of critical points and to Bobrowski and Mukherjee (2015) for their adaptation to point clouds sampled on manifolds.

- $\Lambda \to \infty$ is called the supercritical regime. In this case,

$$\frac{1}{n}\mathbb{E}\left[N_k\left((X_i/r_n)_{i=1}^n\right)\right] \underset{n\to\infty}{\to} \frac{1}{(k+1)!}\int_{\left(\mathbb{R}^d\right)^k} h(0, \mathbf{y})e^{-\omega_d R^d(0,\mathbf{y})}d\mathbf{y},$$

  where h and R are geometric functions that do not depend on f and ω_d is the volume of the unit ball in $\mathbb{R}^d$. Although a similar result has never been derived for Betti numbers,

we can very much expect a similar behavior and that the density of the sampling cannot be extracted from the persistence diagram. Intuitively, the supercritical regime kills all the information related to the sampling.

- $\Lambda \to \lambda \in]0, +\infty[$ is called the critical or thermodynamic regime. Under these circumstances,

$$\frac{1}{n}\mathbb{E}\left[N_k\left((X_i/r_n)_{i=1}^n\right)\right] \underset{n\to\infty}{\to} \frac{\lambda}{(k+1)!}\int_{\left(\mathbb{R}^d\right)^{k+1}} f^{k+1}(x)\tilde{h}(0,\mathbf{y})e^{-\lambda\omega_d R^d(0,\mathbf{y})f(x)}d\mathbf{y}dx,$$

where $\tilde{h}$ is a geometric functional that does not depend on f. As opposed to the previous case, the limiting number of critical simplices of dimension k explicitly depends on the distribution f. Furthermore, Yogeshwaran et al. (2017) have derived a concentration inequality and a central limit theorem (CLT) for β_k in this regime. It has been shown by Kahle (2011) that $\mathbb{E}[\beta_k]$ is of order n, but computing the exact limit is still an open question, especially its dependency on f. Note that the speed of convergence of β_k towards its expectation also depends on f. We can also cite the work of Krebs and Polonik (2019) that establishes a CLT for persistent Betti numbers, and of Biscio et al. (2020) who derives a CLT in the vague topology for the 0 and 1-persistence diagrams of a point process in dimension 2 verifying some mild assumptions.

- $\Lambda \to 0$ is called the sub-critical or sparse regime and is the most favorable case of study. To start with, we have that

$$\frac{1}{n^{k+1}r_n^{dk}}\mathbb{E}\left[N_k\left((X_i/r_n)_{i=1}^n\right)\right] \underset{n\to\infty}{\to} \frac{1}{(k+1)!}\int_{\mathbb{R}} f^{k+1}(x)dx\int_{(R^d)^k}\tilde{h}(0,\mathbf{y})d\mathbf{y}.$$

The dependency on f is much simpler than in the critical regime. In addition, a similar rate and dependency over $\int_{\mathbb{R}} f^{k+1}(x)dx$ can be derived for the Betti numbers as shown by Bobrowski and Mukherjee (2015) in the case of manifolds. For this regime, the analysis has been pushed even further by Owada (2021) for the convergence of persistence diagrams. We generalize his results in the following theorem where our contribution is two-fold: the data are now allowed to be sampled from a manifold, and we provide an alternative proof technique that relies on the Hoeffding's decomposition for U-statistics which gives us an upper-bound on the variance of the persistent Betti numbers and a rate of convergence of the persistence diagram towards its limiting measure. Before stating the theorem, we define the function h_r by

$$h_r\left(x_1,\ldots,x_{k+2}\right) = \mathbb{1}\left\{\beta_k\left(\check{\mathrm{C}}\left(\{x_1,\ldots,x_{k+2}\},r\right)\right)=1\right\},$$

and for $0 \le s \le t \le u \le v \le \infty$,

$$H_{s,t,u,v}(\mathbf{x}) = h_t(\mathbf{x})h_u(\mathbf{x}) - h_t(\mathbf{x})h_v(\mathbf{x}) - h_s(\mathbf{x})h_u(\mathbf{x}) + h_s(\mathbf{x})h_v(\mathbf{x}).$$

Theorem 10 *Let M be a closed $\mathcal{C}^2$ manifold with reach $\tau_M \ge \tau_{\min}$, take $(X_i)_{i=1}^n$ an i.i.d. sample drawn from $f\mathbb{1}_M \mathrm{d}Vol_d$ for a L-Lipschitz density f.*

For $k \in [\![0, d-1]\!]$, denote by μ_k the measure on $\Delta^+ = \{(x,y) : 0 \leq x < y \leq \infty\}$ defined on the rectangles $R_{s,t,u,v} = [s,t) \times [u,v)$ by

$$\mu_k(R_{s,t,u,v}) = \frac{\int_M f^{k+2} \mathrm{d}Vol_d}{(k+2)!} \int_{(\mathbb{R}^d)^{k+1}} H_{s,t,u,v}(0, y_1, \ldots, y_{k+1}) dy_1 \ldots dy_{k+1},$$

for $0 < s \leq t \leq u \leq v$. For a sequence r_n, denote by $\xi_{k,n}$ the re-scaled measure defined by

$$\xi_{k,n}(R_{s,t,u,v}) = \frac{Card(r_n R_{s,t,u,v})}{n^{k+2} r_n^{d(k+1)}}.$$

Assume that we are in the sparse divergence regime, i.e. the sequence r_n verifies:

$$n r_n^d \to 0 \text{ and } n^{k+2} r_n^{d(k+1)} \to \infty.$$

For $k \leq d-4$, choose $r_n = n^{-\frac{k+2}{2+d(k+1)}}$. Then for n large enough,

$$\sup_{0<s\leq t\leq u\leq v\leq t^+} \mathbb{E}\left[(\xi_{k,n} - \mu_k)(R_{s,t,u,v})^2\right] \leq C n^{-\frac{2(k+2)}{2+d(k+1)}}.$$

For $d-4 \leq k \leq d$, choose $r_n = n^{-\frac{k+4}{d(k+3)}}$. Then for n large enough,

$$\sup_{0<s\leq t\leq u\leq v\leq t^+} \mathbb{E}\left[(\xi_{k,n} - \mu_k)(R_{s,t,u,v})^2\right] \leq C n^{-\frac{2}{k+3}},$$

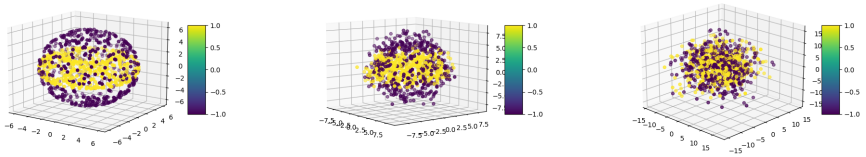
where C is a constant that depends only on $k, d, t^+, \|f\|_\infty, \tau_{\min}$ and L.

We defer the proof to Section 5.7.

This theorem asserts that asymptotically, the rescaled persistence diagram of the Cech filtration built on an adequately rescaled point cloud on $\mathbb{R}^d$ converges to a measure μ_k which depends on the point cloud only through $\int_M f^{k+2} \mathrm{d}Vol_d$. Moreover, given two distributions f_1 and f_2 such that there exists $k \in [\![0, d-1]\!]$ such that $\int_M f_1^{k+2} \mathrm{d}Vol_d \neq \int_M f_2^{k+2} \mathrm{d}Vol_d$, any rectangle $R_{s,t,u,v}$ enables us to distinguish between the two densities f_1 and f_2 when n is large enough as we make it more explicit in the following corollary. Since this theorem is stated for the rescaled persistence diagram with a sequence r_n that tends to 0, this is another evidence that points close to the diagonal (even close to the origin) contain information relative to the sampling and should be considered for classification purposes.

Corollary 11 *Keeping the same notation as above, consider two densities f_1 and f_2 with Lipschitz constants L_1 and L_2 such that there exists $k \in [\![0, d-1]\!]$ such that $\int_M f_1^{k+2} \mathrm{d}Vol_d \neq \int_M f_2^{k+2} \mathrm{d}Vol_d$. Let $0 < s \leq t \leq u \leq v$. For n large enough, the rectangle $R_{s,t,u,v}$ identifies the correct model with probability larger or equal than $1 - C \frac{n^{-\frac{2(k+2)}{d(k+1)}}}{\left|\int_M (f_1^{k+2} - f_2^{k+2})\right|^2}$, where C is a constant that depends only on $(s,t,u,v), k, d, \|f_1\|_\infty, \|f_2\|_\infty, \tau_{\min}, L_1$ and L_2.*

The proof of Corollary 11 is a straightforward consequence of the Chebyshev's inequality and we defer it to Section 5.8. Deriving a finer concentration inequality is still an open



(a) $\sigma = 0$ (b) $\sigma = 1$ (c) $\sigma = 4$

Figure 2: Data to classify. Yellow: torus, purple: sphere.

question: indeed, we have only used a bound on the variance of the random variable $\xi_{k,n}$ to use Chebyshev's inequality. While our proof could be adapted to control higher order moments, it would be worth investigating if we could adapt some techniques from the proof of the Theorem 4.5 of Yogeshwaran et al. (2017) to our framework in the sparse regime.

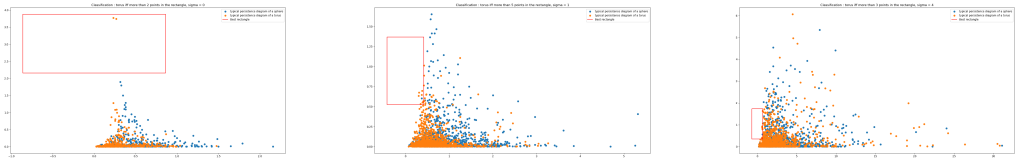
The results derived in Proposition 9 and Corollary 11 both state that the number of sampled points n must be large enough to discriminate between the two sampling models with large probability, whether we want to distinguish between manifolds with different homology or different samplings on the same manifold. On the contrary, the results from the previous sections, especially the dependency over m in Proposition 1 and $\bar{M}_2$ in Theorem 3 and Proposition 5 assert that the number of points in the diagram (directly related to the number of sampled points) must not be too large in order to obtain a good control of the Rademacher complexity. The number of sampled points n acts as a trade-off between the separation of the two classes and a control of the predictive risk.

3.3 Examples

In this section, we allow ourselves to rotate the diagrams by applying the transformation $(x, y) \mapsto (x, y - x)$, so that all the points lie in the upper-right quadrant, and the diagonal is mapped to the x-axis. In order to illustrate our method, we start by considering $n = 500$ points lying on a torus (class $+1$) or a sphere (class 0) and classify it based on the 1-persistence diagram of its Cech complex. The persistence diagram of the torus is expected to have two high-persistence features. Some examples of data are shown in Figure 2 and rectangle classifiers on Figure 3. The sphere has radius 6, and the inner circle of the torus has size 2 while the outer one has size 4.

In the noise-free setting, it is very easy to distinguish between the two classes, both on the raw input and on the persistence diagrams. This corresponds to the framework of Theorem 9. If we add a Gaussian noise, it is no longer possible to distinguish which shape is a torus and which is a sphere based on their homology, but it is still possible to distinguish between them because they have different volume measures, by investigating early-born features.

On another experimental set-up, we still aim at distinguishing between point clouds sampled from a torus or a sphere, except that the size of the supporting manifold as well as the number of points are drawn at random. In addition we add a small isotropic noise to the input sample. The illustration of Figure 4 shows the first four rectangles of the boosting procedure.



(a) $\sigma = 0$ (b) $\sigma = 1$ (c) $\sigma = 4$

Figure 3: Best rectangle to classify points from a sphere or a torus.

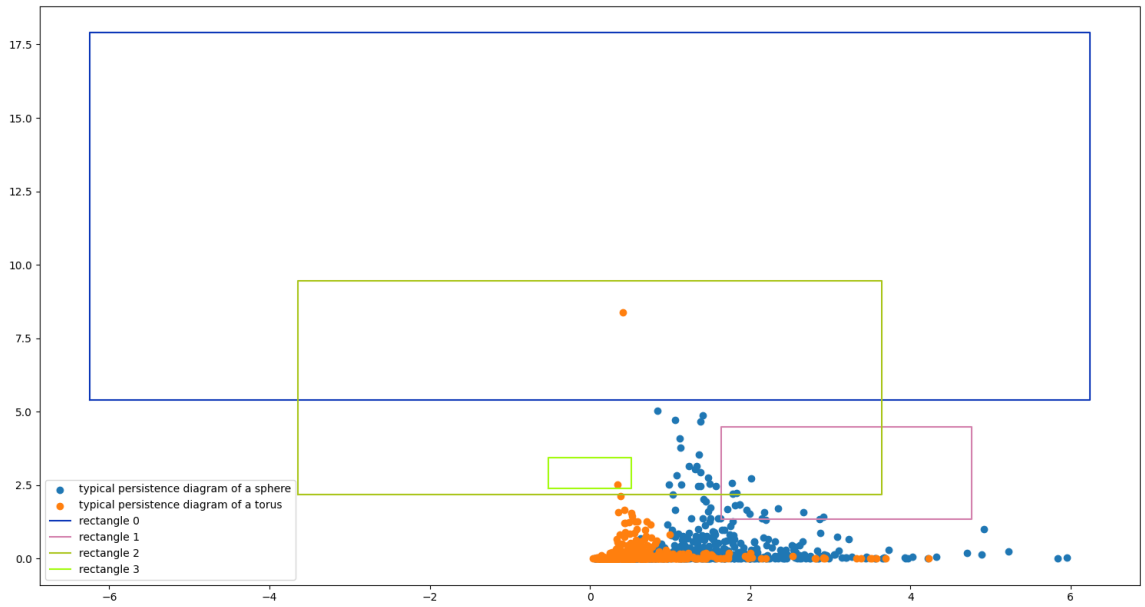
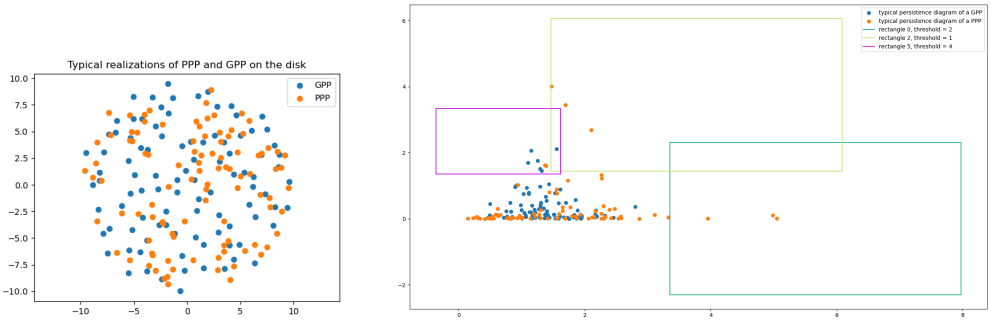


Figure 4: Boosting for manifold classification.

The first rectangle aims at discriminating based on the presence of a high-persistence point in the diagram, that would have it classified as a torus (here, there is only one point because of the added noise that makes one of the two features collapse). In this figure, this rectangle alone would suffice to tell the two data apart. However, on other realizations, some of the topological noise from the sphere also belongs to this rectangle. The second rectangle therefore aims at classifying based on the topological noise. Indeed, for points sampled on a torus, cycles will typically be born earlier than on the sphere, and this rectangle aims at detecting late-born cycles, under which circumstances the data will be classified as "sphere". The third rectangle aims at detecting whether there is a significant number of points of high persistence in the topological noise. The fourth one explores if there are features born early, which is a signature of tori. The boosting algorithm aggregates these classifiers and improves the classification performance by up to 10 % as opposed to considering a single rectangle.

A second experiment conducted is based on the experimental set-up from Obayashi et al. (2018). We sample Poisson (PPP) and Ginibre (GPP) point processes on the disk, with 30 points on average and compute their one-dimensional persistence diagrams? The



(a) Samples of point processes

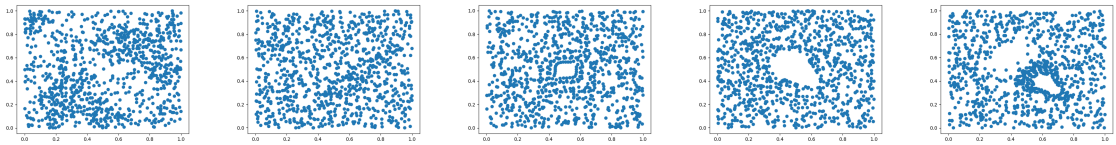
(b) Boosting rectangles

Figure 5: Point processes classification problem.

model has been trained on 400 processes and tested on 200. We have reached similar classification accuracy (around 94% in both cases). They apply a logistic regression to a persistence image transform of the persistence diagrams. When using a L^1 penalty, this induces sparsity and highlights a zone of the persistence image useful for discrimination. Our method can be seen as a variation of this where we are free from vectorization and fixed-pixelization when selecting the discriminating support. It is no surprise we obtain similar results on this simple data set. We will actually see in Section 4 that our method has a better accuracy on real data sets for a comparable running time. We display the results of boosting when 100 points for each process are sampled in Figure 5. A Ginibre point process causes repulsive interactions and points are more evenly spread out, which prevents cycles from dying too early and promotes features with medium-persistence, as we can see on Figure 5. In this set-up, there is no "homological signal" to recover, and we only classify based on the topological noise. We only display three rectangles because of overlaps. The first rectangle investigates very late-born cycles of small persistence, which seems to be a characteristic of PPP. Another rectangle looks at features of high-persistence born late, which is once again something promoted by PPP. On this example, this rectangle alone would bring a misclassification. The last rectangle seeks for features of medium persistence born early, and classify as a GPP if there are more than four such features (which is the case here).

4. Quantitative experiments

We wish to compare our method to benchmark datasets in both topological data analysis and point clouds classification. For all the experiments, we typically perform 10 to 20 boosting iterations where the weak-classifiers are Euclidean balls along with a threshold and where all the parameters are learned by exhaustive search (Algorithm 1 of Section 2.3). At each boosting step, we search for centers of balls among a sub-sample of the k-means clusters' centers. When the number of data is somewhat large, we allow ourselves to optimize only over some subset of the available data, taking new data at each boosting step. The $1/\sqrt{N}$ bounds obtained in Section 2.2 warrant for the validity of this sub-



(a) $\rho = 2.5$ (b) $\rho = 3.5$ (c) $\rho = 4.0$ (d) $\rho = 4.1$ (e) $\rho = 4.3$

Figure 6: Examples of point clouds from the Orbit5K dataset.

PSS-K	PWG-K	SW-K	PF-K	Perslay	Persformer	BBA
72.38 ± 2.4	76.63 ± 0.7	83.6 ± 0.9	85.9 ± 0.8	87.7 ± 1.0	91.2 ± 0.8	83.3 ± 0.5

Table 1: Classification score for the Orbit5K dataset.

sampling procedure. In the tables below, our method will be denoted by BBA for "best balls aggregator". We have made the code publicly available here.[1]

4.1 Persistence diagrams

ORBIT5K

The dataset ORBIT5K is often used as a standard benchmark for classification methods in TDA. This dataset consists of subsets of size 1000 of the unit cube $[0, 1]^2$ generated by a dynamical system that depends on a parameter $\rho > 0$. To generate a point cloud, a random initial point (x_0, y_0) is chosen uniformly in $[0, 1]^2$ and a sequence of points (x_n, y_n) for $n = 0, 1, \ldots, 999$ is generated recursively by:

$$\begin{aligned} x_{n+1} &= x_n + \rho y_n (1 - y_n) \mod 1 \\ y_{n+1} &= y_n + \rho x_{n+1} (1 - x_{n+1}) \mod 1. \end{aligned}$$

Given an orbit, we want to predict the value of ρ, that can take values in $\{2.5, 3.5, 4.0, 4.1, 4.3\}$. We display an example for each class in Figure 6; $\rho \in \{4.0, 4.1, 4.3\}$ accounts for difference in topology, while $\rho \in \{2.5, 3.5\}$ generates samplings with different densities but no particular homological information.

We generate 700 training and 300 testing data for each class. We perform a one-versus-one classification. We compare our score with standard classification methods in Table 1, where the results are averaged over 10 runs. We compare our scores to four kernel methods on persistence diagrams taken respectively from Reininghaus et al. (2015), Kusano et al. (2016), Carriere et al. (2017), Le and Yamada (2018), and two methods that use a neural network architecture to vectorize persistence diagrams: Carrière et al. (2020) and Reinauer et al. (2021). Our accuracy is comparable with kernel methods on persistence diagrams but is somehow lower than that of neural networks.

GRAPH CLASSIFICATION

Another benchmark of experiments in TDA is the classification of graph data. In order to transform graphs into persistence diagrams, we consider the Heat Kernel Signature (HKS) as

[1] `https://github.com/OlympioH/BBA_measures_classification`

Dataset	SV	RetGK	FGSD	GCNN	GIN	Perslay	P	MP	Atol	BBA
Mutag	88.3	90.3	92.1	86.7	89	89.8	79.2	86.1	88.3	90.4
DHFR	78.4	81.5	-	-	-	80.3	70.9	81.7	82.7	80.5
Proteins	72.6	75.8	73.4	76.3	74.8	75.9	65.4	67.5	71.4	74.7
Cox2	78.4	80.1	-	-	-	80.9	76.0	79.9	79.4	81.2
IMDB-B	72.9	71.9	73.6	73.1	74.3	71.2	54.0	68.7	74.8	69.4
IMDB-M	50.3	47.7	52.4	50.3	52.1	48.8	36.3	46.9	47.8	46.7
COLLAB	-	81.0	80.0	79.6	80.1	76.4	-	-	88.3	69.6

Table 2: Classification scores for graph data.

done by Carrière et al. (2020), for which we recall the construction: for a graph $\mathcal{G} = (V, E)$, the HKS function with diffusion parameter t is defined for each $v \in V$ by

$$hks_t(v) = \sum_{k=1}^{|V|} \exp(-t\lambda_k)\psi_k(v)^2,$$

where λ_k is the k-th eigenvalue of the normalized graph Laplacian and ψ_k the corresponding eigenfunction. We build two persistence diagrams of dimensions 0 and 1 tracking the evolution of the topology of the sublevel sets of this function, and kept whichever one gave the best results. For the experiments, we fixed the value of t to 10, a preliminary study suggested that the diagrams were somehow robust to the choice of this diffusion parameter. The results on standard datasets are provided in Table 2. The first five columns are kernel methods or neural networks designed specifically for graph data, P denotes the best method between Persistence image and Persistence landscapes, and MP the best method between multiparameter persistence image, landscape, and kernel (scores reported from Carrière and Blumberg (2020)). All these persistence-based vectorizations are coupled with a XGBoost classifier to perform the learning task. We can see that our method clearly outperforms standard vectorizations of persistence diagrams and also multi-persistence descriptors. The accuracy reached is similar to Perslay, Carrière et al. (2020) which is a neural network that learns a vector representation of a persistence diagrams and Atol, Royer et al. (2021) which is another measure learning method. Note that on the biggest dataset COLLAB, our method is clearly outperformed by the other methods, especially Atol.

4.2 Other datasets

FLOW CYTOMETRY

Flow cytometry is a lab test used to analyze cells' characteristics. It is used to perform a medical diagnosis by measuring various biological markers for each cell in the sample. Mathematically, the data are point clouds consisting of tens of thousands of cells living in $\mathbb{R}^D$, where D is the number of biological markers considered. We have trained our model on the Acute Myeloid Leukemia (AML) dataset available here[1]. AML is a type of blood cancer that can be detected by performing flow cytometry on the bone marrow cells. The dataset consists of 359 patients, half of them are used for training and the rest of them for

[1] `https://flowrepository.org/id/FR-FCM-ZZYA`

Dataset	dimension	classes	CMFM	LCEM	XGBM	RFM	MLSTM-FCN	ED	DTW	BBA
Heartbeat	61	2	76.8	76.1	69.3	80	71.4	62	71.7	73.7
SCP1	6	2	82	83.9	82.9	82.6	86.7	77.1	77.5	77.5
SCP2	7	2	48.3	55.0	48.3	47.8	52.2	48.3	53.9	56.0
Finger Movements	28	2	50.1	59.0	53.0	56.0	61.0	55.0	53.0	58.0
Epilepsy	3	4	99.9	98.6	97.8	98.6	96.4	66.7	97.8	92.8
StandWalkJump	4	3	36.3	40	33.3	46.7	46.7	20	33.3	46.7
Racket Sports	6	4	80.9	94.1	92.8	92.1	88.2	86.8	84.2	73.7

Table 3: Classification scores for multi-dimensional time series dataset.

validating the model. For each patient, 7 biological markers are measured across 30000 cells. We report a $F1$-score of 98.9 %, while most flow cytometry specific data analysis methods have a score comprised between 95% and 100% according to Table 3 from Aghaeepour et al. (2013). In addition, our method can lead to qualitative interpretations, since it generates discriminatory zones, and therefore thresholds of activation for biological markers that make a patient sick or healthy.

Time series

Another field of applications is the classification of time series. We consider each data as a collection of points by dropping the temporal aspect of the data. We have tried our method on a small sample of data from the University of East Anglia (UEA) archive presented in Bagnall et al. (2018). We compare our method against standard classification methods, and report the results from Baldán and Benítez (2021) in Table 3

Our method competes with the most simple methods for classifying time series, but fails to be state of the art, especially when a high classification score is expected. When there is only little information to be captured (for instance for the datasets StandWalkJump, Finger Movements or SCP2), our method manages to retrieve it. It is to be noted that the comparison cannot be completely fair with respect to methods targeted to specifically deal with time series while we have removed the temporal aspect of the data and only focus on the distribution of the d-dimensional data in certain areas of $\mathbb{R}^d$.

4.3 Discussion

Computational time

In order to compare the running time of our method with standard vectorization methods, we consider the problem defined in Section 3.3: we observe points on a torus or a sphere and classify the manifold supporting the point clouds based on their one-dimensional persistence diagrams. We assume the diagrams have been computed in a preliminary step and compare the running time of several methods in Table 4 when classifying over a training set of size 500 or 3000. The average number of points in the one-dimensional persistence diagram is experimentally of the same order as the number of sampled points. For our method, we report the training time for one weak-classifier. In this experiment only one weak-classifier is enough to classify. For the exhaustive search, we have looked for a candidate classifier among a family of balls with 20 different centers, 10 different radii and 5 different thresholds for a total of 2000 possible classifiers, counting reversed labels. We compare the running times with a Persistence Image of resolution 40×40 with fixed parameters and we train a

Table 4: Computational time (in seconds), torus VS sphere.

Size of the dataset	Size of the point cloud	BBA (smooth)	BBA (exhaustive search)	Lasso-PI + logit-L2
500	100	143.0	18.7	7.4
3000	100	139.5	97.6	151.4
500	500	136.8	31.8	12.0
3000	500	139.2	180.1	125.7
500	2000	147.4	66.6	43.1
3000	2000	149.7	343.9	260.1

Table 5: Computational time (in seconds), graph data

Name of the dataset	Number of data	BBA (smooth)	BBA (exhaustive search)	Lasso-PI + logit-L2	Atol (vectorization only)
Mutag	170	40.8	3.0	0.27	< 0.1
Collab	4500	195.1	172.5	164.9	110
Collab	1000	175	38.6	36.2	-

logit classifier with a L^2 penalty, where the regularization parameter is learned by cross-validation. When the number of sampled points is large enough, most of the computation time is devoted to the vectorization part and only a small fraction of it is dedicated to actually classifying the images. When the number of points is too small, the classification part of the pipeline can take a rather long time.

The implementation of vectorization methods for persistence diagrams and standard classification algorithms are taken respectively from the Gudhi (Maria et al. (2014)) and Scikit-learn (Pedregosa et al. (2011)) libraries. It is likely that our implementation can be improved, leading to a potential computational gain. Nevertheless, an exhaustive search of the best ball-classifier has a comparable running time to that of Lasso-PI + logit L2 which is enough for simple examples. When doing an aggregation procedure of several weak-classifiers, the running time becomes significantly longer but provides a greater accuracy, as noted in Table 2. It is also to be noted from Table 4 that our implementation of the optimization of the smoothed objective does not vary much when dealing with large point clouds nor with large datasets, which makes it a preferable candidate for large-scale applications.

This is backed-up by the timing of some of the graph experiments in Table 5 where we also compare our running times with the Atol method from Chazal et al. (2021) for the smallest and biggest graph datasets. For the Atol method, the authors only report the vectorization time without taking into account the training time of a random forest. Note that the average number of nodes and edges in the Mutag dataset are 17.9 and 19.8 while they are of 74.5 and 2457.2 for the Collab dataset, and our method seems to be pretty robust in this increase in scale. We can see that the running time of all methods is comparable. However, in our case, the accuracy of a single weak classifier is quite poor and the BBA method requires about 10 boosting steps to be fully competitive. For small datasets, both in terms of number of points and data, an exhaustive search is highly recommended, also due to the unstable nature of the smooth version which often requires several initializations before finding a relevant classifier.

Take-home message

The method developed in this article, while being simple and explainable, allows to tackle a wide variety of problems. When used on persistence diagrams, we obtain similar results as kernel methods and manage to come close to some state-of-the-art methods using neural networks on graph data. Our method has a decent performance in terms of accuracy when used on small datasets. When the number of data is larger, the $1/\sqrt{N}$ bounds from Section 2.2 justify for training our model on a sub-sample of the dataset and therefore propose a decent accuracy at a mild computational cost.

In addition, since we locate the areas of the persistence diagram which are the most relevant for classification, this can give information for truncating the simplicial complexes for future applications on the same type of data, and therefore greatly improve the computational time, especially if one is to compute the Rips complex which is known to have a prohibitive number of simplices if untruncated. Due to its simplicity, the natural competitors of our method appear to be standard vectorizations of persistence diagrams coupled with a usual learning algorithm such as logit or random forest. For this type of classifier, we have seen in Table 2 that our method has a greater accuracy, while having a comparable running time. Beyond persistence diagrams, we have demonstrated that our method offers decent results in a variety of settings and is well suited to dealing with simple data and could be adapted to dealing with large-scale applications.

5. Proofs

This section is devoted to the proofs of all theoretical results contained throughout the paper.

5.1 Proof of Proposition 1

We denote by $\tilde{\mathcal{F}}$ the class of functions on measures defined by $\tilde{f}(\mu) = [f(x_1), \dots, f(x_n)]$ for $\mu = \sum_{i=1}^n \delta_{x_i} \in \mathcal{M}_m(\mathcal{X})$ and $f \in \mathcal{F}$. We denote by $k \mapsto \gamma_{\mathcal{F}}(k)$ the growth function of a hypothesis class $\mathcal{F}$ defined by $\gamma_{\mathcal{F}}(k) = \sup_{x_1,\dots,x_k} \sharp\{(f(x_1), \dots, f(x_k) | f \in \mathcal{F}\}$. We have that $\gamma_{\tilde{\mathcal{F}}}(N) \leq \gamma_{\mathcal{F}}(mN)$ since all the measures have at most m points. Using the Sauer-Shelah lemma, we therefore have that $\gamma_{\tilde{\mathcal{F}}}(N) \leq \left(\frac{emN}{d}\right)^d$ where d is the VC-dimension of $\mathcal{F}$.

Now, consider a set of d_2 measures that is shattered by the class $\mathcal{H}$. Using Section 20 of Shalev-Shwartz and Ben-David (2014), we have that for every integer k, $\gamma_{\mathcal{H}}(k) \leq \gamma_{\tilde{\mathcal{F}}}(k)\gamma_{\tilde{\mathcal{G}}}(k)$ using the observation above Proposition 1 that $\mathcal{H}$ is a composition class. We therefore have, using Sauer-Shelah lemma again, that:

$$2^{d_2} \leq \gamma_{\mathcal{H}}(d_2) \leq \left(\frac{emd_2}{d}\right)^d \left(\frac{ed_2}{d'}\right)^{d'}.$$

Taking the logarithm on both sides and using the same computation as in the proof of Theorem 6 of Sabato and Tishby (2012) yields the wanted result.

5.2 Proof of Lemma 2

Let h and g be two functions of $\mathcal{F}$.

$$\|\tilde{g}-\tilde{h}\|_{L_p^N}=\left(\frac{1}{N}\sum_{i=1}^{N}(g[\mu_i]-h[\mu_i])^p\right)^{1/p}$$

$$\left|\int(h-g)\mathrm{d}\mu_i\right|^p=M_i^p\left|\int(h-g)\mathrm{d}(\mu_i/M_i)\right|^p$$
$$\leq M_i^p\int(h-g)^p\mathrm{d}(\mu_i/M_i) \text{ by Jensen inequality.}$$

Therefore,

$$\|\tilde{h}-\tilde{g}\|_{L_p^N}^p\leq\frac{1}{N}\sum_{i=1}^{N}M_i^p\int(h-g)^p\mathrm{d}(\mu_i/M_i).$$

Denoting for each i $w_i=\frac{M_i^p}{\sum_{j=1}^N M_j^p}$, the above inequality writes as :

$$\begin{aligned}\|\tilde{h}-\tilde{g}\|_{L_p^N}^p&\leq\bar{M}_p\sum_{i=1}^{N}w_i\int(h-g)^p\mathrm{d}(\mu_i/M_i)\\&\leq\bar{M}_p\|h-g\|^p_{L_p(\sum_{i=1}^N w_i\mu_i/M_i)}.\end{aligned}$$

Denoting by $\bar{\mu}=\sum_{i=1}^N w_i\mu_i/M_i$ we have the desired result.

5.3 Proof of Theorem 3

By Dudley's chaining theorem, we have that

$$\mathcal{R}_N(\tilde{\mathcal{F}})\leq\frac{12}{\sqrt{N}}\int_0^\infty\sqrt{\ln\mathcal{N}(\tilde{\mathcal{F}},L_2^n,\varepsilon)}d\varepsilon.$$

Remark that

$$\mathrm{diam}(\tilde{\mathcal{F}},L_2^n)\leq\frac{1}{\sqrt{N}}\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{N}\int(f\mathrm{d}\mu_i)^2\right)^{1/2}\leq\frac{1}{\sqrt{N}}\left(\sum_{i=1}^{N}M_i^2\right)^{1/2}\leq\bar{M}_2.$$

Therefore, we only need to integrate up to $\bar{M}_2$, yielding:

$$
\begin{aligned}
\mathcal{R}_N(\tilde{\mathcal{F}}) &\leq \frac{12}{\sqrt{N}} \int_0^{\bar{M}_2} \sqrt{\ln \mathcal{N}(\mathcal{F}, L_2(\bar{\mu}), \varepsilon/\bar{M}_2)} d\varepsilon \text{ by the above lemma,} \\
&\leq \frac{12}{\sqrt{N}} \int_0^{\bar{M}_2} \sqrt{K_0 vc(\mathcal{F}, c\varepsilon/\bar{M}_2) \ln(2\bar{M}_2/\varepsilon)} d\varepsilon \text{ by Theorem 1 of Mendelson and Vershynin (2003),} \\
&\leq \frac{12\bar{M}_2}{\sqrt{N}} \int_0^1 \sqrt{K_0 vc(\mathcal{F}, c\varepsilon) \ln(2/\varepsilon)} d\varepsilon \text{by a change of variables,} \\
&\leq \frac{K_1 \bar{M}_2 \sqrt{vc(\mathcal{F})}}{\sqrt{N}} \int_0^1 \sqrt{\ln(2/\varepsilon)} d\varepsilon.
\end{aligned}
$$

Here, K_0 and K_1 are universal constants. Including the integral in the multiplicative constant term gives the wanted result.

5.4 Proof of Theorem 4

By Sudakov minoration principle, there exists a constant C such that for all $\varepsilon > 0$,

$$
\frac{C\varepsilon}{\sqrt{N}} \sqrt{\ln \mathcal{N}(\tilde{\mathcal{F}}, \varepsilon, L_2^N)} \leq G_N(\tilde{\mathcal{F}}),
$$

where G_N stands for the Gaussian complexity.

Classical equivalence between covering and packing numbers yields

$$
\frac{C\varepsilon}{\sqrt{N}} \sqrt{\ln \mathcal{M}(\tilde{\mathcal{F}}, 2\varepsilon, L_2^N)} \leq G_N(\tilde{\mathcal{F}}).
$$

If all the μ_i are of the form $M_i \delta_{x_i}$ for $(x_i)_{i=1,\dots,N} \in \mathcal{X}^N$, we have for two functions g and h in $\mathcal{F}$ that

$$
\|\tilde{g} - \tilde{h}\|_{L_2(\mu_1^N)} = \frac{1}{N} \sqrt{\sum_{i=1}^N (g[\mu_i] - h[\mu_i])^2} = \frac{1}{N} \sqrt{\sum_{i=1}^N M_i^2 (g(x_i) - h(x_i))^2} = \bar{M}_2 \|g - h\|_{L_2(x_1^N, w)},
$$

for the L_2-norm with weights $w_i = \frac{M_i^2}{\sum_{j=1}^N M_i^2}$.

When looking on the supremum over all measures, we can therefore lower bound the packing number:

$$
\begin{aligned}
\sup_{(\mu_1,\dots,\mu_N) \in \mathcal{C}_{\bar{M}_2}^N} G_N(\tilde{\mathcal{F}}) &\geq \frac{C\varepsilon}{\sqrt{N}} \ln \sqrt{\sup_{(\mu_1,\dots,\mu_N) \in \mathcal{C}_{\bar{M}_2}^N} \mathcal{M}(\tilde{\mathcal{F}}, 2\varepsilon, L_2(\mu_1^N))} \\
&\geq \frac{C\varepsilon}{\sqrt{N}} \ln \sqrt{\sup_{x_1,\dots,x_N} \mathcal{M}(\mathcal{F}, 2\varepsilon/\bar{M}_2, L_2(x_1^N, w))}.
\end{aligned}
$$

In particular, by taking $x_1, \dots x_N$ that are $2\varepsilon/\bar{M}_2$-shattered by $\mathcal{F}$ if $N \leq vc(\mathcal{F}, 2\varepsilon/\bar{M}_2)$ along with uniform weights, Proposition 1.4 from Talagrand (2003) states that the logarithm

of the packing number dominates the fat-shattering function. If $N > vc(\mathcal{F}, 2\varepsilon/\bar{M}_2)$, the same result simply follows by considering the uniform measure on $vc(\mathcal{F}, 2\varepsilon/\bar{M}_2)$ of the N points and setting weight 0 to the others.

This together with the equivalence between Gaussian and Rademacher complexities yields that for all $\varepsilon > 0$,

$$K' \frac{\varepsilon}{\sqrt{N}\ln(N)} \sqrt{vc(\mathcal{F}, 4\varepsilon/\bar{M}_2)} \leq \sup_{(\mu_1,\dots,\mu_N)\in\mathcal{C}^N_{\bar{M}_2}} \mathcal{R}_N\left(\tilde{\mathcal{F}}|\mu_1,\dots,\mu_N\right).$$

In particular, taking $\varepsilon = \bar{M}_2/8$ gives the wanted result, by noticing that for the classification problem, i.e. labels in $\{0,1\}$, we have that $vc(\mathcal{F}, 1/2) = vc(\mathcal{F})$.

5.5 Proof of Proposition 5

Let $\bar{\mu} = (\mu_1, \dots, \mu_N)$ be a sample of N measures. In Theorem 2 of Maurer (2016), the authors establish a chain-rule to control the Gaussian complexity for the composition of function classes. This result implies that there exist two constants C_1 and C_2 such that for any $f_0 \in \mathcal{F}$,

$$G_N(\mathcal{H}) \leq C_1 L G_N(\tilde{\mathcal{F}}) + \frac{1}{N} C_2 \mathrm{Diam}(\tilde{\mathcal{F}}(\bar{\mu}))\mathbf{R}(\mathcal{G}) + G_N(\mathcal{G}(f_0)).$$

$$\text{where } \mathbf{R}(\mathcal{G}) = \sup_{\mathbf{x},\mathbf{x}'\in\mathbb{R},\mathbf{x}\neq\mathbf{x}'} \mathbb{E}_\gamma \sup_{\psi\in\mathcal{H}} \frac{(\psi(\mathbf{x}) - \psi(\mathbf{x}'))\gamma}{|\mathbf{x} - \mathbf{x}'|},$$

and where $\gamma \sim \mathcal{N}(0,1)$.

We wish to successively bound each of the three terms on the right hand side. The classical equivalence between Gaussian and Rademacher complexities together with Theorem 3 permits to control the first term:

$$G_N(\mathcal{H}) \leq \frac{C_1 L \bar{M}_2 \sqrt{vc(\mathcal{F})}\sqrt{\log(N)}}{\sqrt{N}} + \frac{1}{N} C_2 \mathrm{Diam}(\tilde{\mathcal{F}}(\bar{\mu}))\mathbf{R}(\mathcal{G}) + G_N(\mathcal{G}(f_0)).$$

Analogously to the proof of Theorem 3, we can simply bound the diameter by $\mathrm{Diam}(\tilde{\mathcal{F}}(\bar{\mu})) \leq \sqrt{N}\bar{M}_2$, since all the functions from $\mathcal{F}$ are bounded by 1.

As for the third term, taking $f_0 = 0$, we have that

$$\begin{aligned} G_N(\mathcal{G}(\tilde{f}_0) &= \mathbb{E}_\gamma \left[\sup_{\psi\in\mathcal{H}} \langle \gamma, (\psi(0), \dots, \psi(0)\rangle \right] \\ &\leq \mathbb{E}_\gamma \left[\left| \sum_{i=1}^N \gamma_i \right| \times \sup_{\psi\in\mathcal{H}} |\psi(0)| \right] \\ &\leq \sqrt{N} \sup_{\psi\in\mathcal{H}} |\psi(0)|, \end{aligned}$$

where the last inequality follows from the fact that the γ_i are standard independent normal variables.

5.6 Proof of Theorem 9

Assume that we observe a n-sample from $\mathbb{M}_1$. Corollary 3 of Chazal et al. (2014b) states that for every $\varepsilon > 0$,

$$\mathbb{P}[d_B(dgm(\check{\mathrm{C}}(\mathbb{M}_1)), dgm(\check{\mathrm{C}}(\hat{X}_n))) \geq \varepsilon] \leq \frac{2^b}{a\varepsilon^b} \exp(-na\varepsilon^b).$$

Let $\delta > 0$, and take $\varepsilon = K/2$. The above formula yields that for $n \geq \frac{2^b}{aK^b} \log\left(\frac{4^b}{aK^b\delta}\right)$, we have with probability larger than $1-\delta$ that

$$d_B(dgm(\check{\mathrm{C}}(\mathbb{M}_1)), dgm(\check{\mathrm{C}}(\hat{X}_n))) \leq K/2. \tag{1}$$

By triangle inequality for the distance d_B and using the hypothesis that the persistence diagrams of the two metric spaces are away from at least K for the bottleneck distance, we necessarily have that

$$d_B(dgm(\check{\mathrm{C}}(\mathbb{M}_2)), dgm(\check{\mathrm{C}}(\hat{X}_n))) > K/2. \tag{2}$$

By assumption, for $i = 1, 2, dgm(M_i)$ has no point below K. We can now distinguish two cases :

- $dgm(\check{\mathrm{C}}(\mathbb{M}_1))$ and $dgm(\check{\mathrm{C}}(\mathbb{M}_2))$ have the same number of points m (all these points are at least away from K to the diagonal). Under these circumstances, $dgm(\check{\mathrm{C}}(\hat{X}_n))$ also has m points above $K/2$. If it had more, it would mean that one of this point should be matched with the diagonal, and therefore yields a contradiction with (1). Consider squares of size $K/2$ centered on the points of $dgm(\check{\mathrm{C}}(\mathbb{M}_1))$ and $dgm(\check{\mathrm{C}}(\mathbb{M}_2))$. If they all contain the same number of points from $dgm(\check{\mathrm{C}}(\hat{X}_n))$, we have a contradiction with (2). It therefore means that there is a rectangle that can select the right model.
- If they do not have the same number of points, necessarily by (1), $dgm(\check{\mathrm{C}}(\hat{X}_n))$ must have the same number of points as $dgm(\check{\mathrm{C}}(\mathbb{M}_1))$ above $K/2$ and do not have the same number of points as $dgm(\check{\mathrm{C}}(\mathbb{M}_2))$. Counting the number of points in the (infinite but truncatable) rectangle $\{(p,q) | \|(p,q)\|_\infty > K/2\}$ is therefore enough to classify between the two metric spaces.

5.7 Proof of Theorem 10

First note that $\mathrm{Card}(r_n R_{s,t,u,v}) = \beta_{k,n}(t,u) - \beta_{k,n}(t,v) - \beta_{k,n}(s,u) + \beta_{k,n}(s,v)$, where $\beta_{k,n}(s,t)$ is the persistent Betti number of order k between times s and t, that is the number of k-cycles that appear in $\check{\mathrm{C}}(r_n^{-1}\mathbb{X}_n)$ before time s and are still alive at time t. As in Owada (2021), we denote by

$$h_r(x_1, \dots x_{k+2}) = \mathbb{1}_{\bigcap_{j_0=1}^{k} \{\bigcap_{j \neq j_0} \mathcal{B}(x_j, r/2) \neq \emptyset\}} - \mathbb{1}_{\bigcap_{j=1}^{k+2} \mathcal{B}(x_j, r/2) \neq \emptyset},$$

and by

$$G_{k,n}(s,t) = \sum_{\mathcal{Y} \subset \mathbb{X}_n, |\mathcal{Y}| = k+2} h_{r_n s}(\mathcal{Y}) h_{r_n t}(\mathcal{Y}),$$

so that, according to (Owada, 2021, Lemma 4.1)

$$G_{k,n}(s,t) - \binom{k+3}{k+2} L_{r_n t} \leq \beta_{k,n}(s,t) \leq G_{k,n}(s,t) + \binom{k+3}{k+1} L_{r_n t}, \tag{3}$$

where

$$L_{r_n t} = \sum_{\mathcal{Y} \subset \mathbb{X}_n, |\mathcal{Y}|=k+3} \mathbb{1}_{\check{\mathrm{C}}(\mathcal{Y}, r_n t) \text{ is connected}}.$$

In what follows we prove bounds for $\beta_{k,n}(s,t)$. The bound on $\mathrm{Card}(r_n R_{s,t,u,v})$ easily follows.

Upper-bound of the bias

$$\begin{aligned}
\mathbb{E}(G_{k,n}(s,t)) &= \binom{n}{k+2} \int_M f(x_1) \mathrm{d}Vol_d(x_1) \int_{M^{k+1}} g_{s,t}\left(\frac{x_1}{r_n}, \ldots, \frac{x_{k+2}}{r_n}\right) \prod_{j=2}^{k+1} f(x_j) \mathrm{d}Vol_d(x_j) \\
&= \binom{n}{k+2} \int_M f(x_1) \mathrm{d}Vol_d(x_1) I_{x_1},
\end{aligned}$$

where $g_{s,t} = h_s h_t$. Now, for a fixed $x_1 \in M$, we note that $g_{s,t}$ is non-zero implies $(x_2, \ldots x_{k+1}) \in \mathcal{B}(x_1, r_n(k+2)t^+)^k$ (recall that $t \leq t^+$). Denoting by $\tilde{M}_n = h_n(M)$, with $h_n : u \mapsto \frac{u - x_1}{r_n}$, and using (Federer, 1959, Theorem 3.1) leads to the change of variable

$$\begin{aligned}
I_{x_1} &:= \int_{M^{k+1}} g_{s,t}\left(\frac{x_1}{r_n}, \ldots, \frac{x_{k+2}}{r_n}\right) \prod_{j=2}^{k+1} f(x_j) \mathrm{d}Vol_d(x_j) \\
&= r_n^{d(k+1)} \int_{(\tilde{M}_n)^{k+1}} g_{s,t}\left(0, y_1, \ldots, y_{k+1}\right) \mathbb{1}_{\mathcal{B}(0,(k+2)t^+)^{k+1}}(y_1, \ldots, y_{k+1}) \prod_{j=1}^{k+1} f(x_1 + r_n y_j) \mathrm{d}Vol_d(y_j).
\end{aligned}$$

Note that $0 \in \tilde{M}_n$, and that $\tilde{M}_n$ has a reach $\tilde{\tau} = \tau / r_n \to +\infty$. With a slight abuse of notation, we identify $T_0 \tilde{M}_n$ with $\mathbb{R}^d$, and denote by J_v the Jacobian of the exponential map $\exp_0 : B_{\mathbb{R}^d}(0, (k+2)t^+) \to \tilde{M}_n$ at point v (note that $\exp_0$ is well defined for n large enough so that $\tilde{\tau} \geq 4(k+2)t^+$, see, for instance (Aamari and Levrard, 2019, Lemma 1)). Using (Federer, 1959, Theorem 3.1) again yields for the change of variable $y_j = \exp_0(v_j)$ that

$$I_{x_1} = r_n^{d(k+1)} \int_{(R^d)^{k+1}} g_{s,t}\left(0, y_1, \ldots, y_{k+1}\right) \mathbb{1}_{y_1, \ldots, y_{k+1} \in \mathcal{B}_{\tilde{M}_n}(0,(k+2)t^+)^{k+1}} \prod_{j=1}^{k+1} J_{v_j} f(x_1 + r_n y_j) dv_1 \ldots dv_{k+1}.$$

According to (Aamari and Levrard, 2019, Lemma 1), whenever $y_j \in \mathcal{B}_{\tilde{M}_n}(0, (k+2)t^+)$, we have $\|y_j - v_j\| \leq C((k+2)t^+)^2 r_n / \tau_{\min}$, and $\|\mathrm{d}_{v_i} \exp_0 - I_d\|_{op} \leq \frac{5}{4\tilde{\tau}_n} = \frac{5 r_n}{4\tau} \leq \frac{5 r_n}{4\tau_{\min}}$, so that

$$|J_{v_j} - 1| \leq C_d \frac{r_n}{\tau_{\min}},$$

and therefore, $\left|J_{v_j}-1\right| \leq 1$ for n large enough. We deduce that

$$\begin{aligned}\left|\prod_{j=1}^{k+1} J_{v_j} f(x_1+r_n y_j) - f(x_1)^{k+1}\right| &\leq \left|\prod_{j=1}^{k+1} J_{v_j} f(x_1+r_n y_j) - \prod_{j=1}^{k+1} J_{v_j} f(x_1)\right| \\ &+ \quad \left|\prod_{j=1}^{k+1} J_{v_j} f(x_1) - f(x_1)^{k+1}\right| \\ &\leq C_d^{k+1}(k+1)L\|f\|_\infty^k r_n + (k+1)\|f\|_\infty^{k+1} C_d^{k+1}\frac{r_n}{\tau_{\min}} \\ &\leq C_d^{k+1}(k+1)\|f\|_\infty^k \left(L \vee \frac{\|f\|_\infty}{\tau_{\min}}\right) r_n.\end{aligned}$$

Denoting by

$$I'_{x_1} = r_n^{d(k+1)} f(x_1)^{k+1} \int_{(R^d)^{k+1}} g_{s,t}\left(0, y_1, \ldots, y_{k+1}\right) \mathbb{1}_{y_1,\ldots,y_{k+1}\in \mathcal{B}_{\tilde{M}_n}(0,(k+2)t^+)^{k+1}} \prod_{j=1}^{k+1} \mathrm{d}v_1 \ldots \mathrm{d}v_{k+1},$$

we deduce that

$$\begin{aligned}|I'_{x_1} - I_{x_1}| &\leq (2(k+2)t^+)^{d(k+1)} C_d^{k+1}(k+1)\|f\|_\infty^{k+1}\left(L \vee \frac{1}{\tau_{\min}}\right) r_n^{d(k+1)+1} \\ &\leq C_{d,k}(t^+)^{d(k+1)}\|f\|_\infty^k \left(L \vee \frac{\|f\|_\infty}{\tau_{\min}}\right) r_n^{d(k+1)+1}.\end{aligned}$$

Next, note that

$$g_{s,t}\left(0, y_1, \ldots, y_{k+1}\right) \neq g_{s,t}\left(0, v_1, \ldots, v_{k+1}\right) \Rightarrow (v_1, \ldots, v_{k+1}) \in V_1,$$

where

$$\begin{aligned}V_1 = \Big\{&(v_1, \ldots, v_{k+1}) \in \mathcal{B}(0, 2(k+2)t^+)^{k+1} \mid \exists i \neq j; \left|\|v_i - v_j\| - s\right| \leq C((k+2)t^+)^2 r_n/\tau_{\min} \\ &\text{or} \quad \left|\|v_i - v_j\| - t\right| \leq C((k+2)t^+)^2 r_n/\tau_{\min}\Big\}.\end{aligned}$$

We deduce that

$$\begin{aligned}&\left|I'_{x_1}(A) - r_n^{d(k+1)} f(x_1)^{k+1} \int_{(R^d)^{k+1}} g_{s,t}\left(0, v_1, \ldots, v_{k+1}\right) \mathbb{1}_{v_1,\ldots v_{k+1}\in \mathcal{B}_{\tilde{M}_n}(0,2(k+2)t^+)^{k+1}} \mathrm{d}v_1 \ldots \mathrm{d}v_{k+1}\right| \\ &\leq r_n^{d(k+1)}\|f\|_\infty^{k+1} \int_{(R^d)^{k+1}} \mathbb{1}_{V_1}(v_1, \ldots, v_{k+1}) \mathrm{d}v_1, \ldots, \mathrm{d}v_{k+1} \\ &\leq r_n^{d(k+1)}\|f\|_\infty^{k+1} 2\binom{k+1}{2} C_d((2(k+2)t^+)^{kd}((2(k+2)t^+)^{d+1}\frac{r_n}{\tau_{\min}} \\ &\leq r_n^{d(k+1)} C_{d,k}\|f\|_\infty^k (t^+)^{(k+1)d}\frac{\|f\|_\infty t^+ r_n}{\tau_{\min}}.\end{aligned}$$

The triangle inequality gives

$$\left|\frac{\mathbb{E}(G_{k,n}(s,t))}{\binom{n}{k+2}r_n^{d(k+1)}} - A_k(s,t)\right| \leq C_{d,k,t^+,\|f\|_\infty}\left(L \vee \frac{\|f\|_\infty}{\tau_{\min}}\right) r_n,$$

where $A_k(s,t) = \left(\int_M f^{k+1}(u)\mathcal{H}^d(du)\right)\int_{(R^d)^{k+1}} g_{s,t}\left(0, v_1, \ldots, v_{k+1}\right) \mathrm{d}v_1 \ldots \mathrm{d}v_{k+1}$.

Next, we have to bound the higher order term $\mathbb{E}(L_{r_n,t})$ dealing with the subsets of size $k+3$. To do so, write

$$\begin{aligned}
\mathbb{E}(L_{r_n t}) &= \binom{n}{k+3}\int_{M^{k+3}} \mathbb{1}_{\check{\mathrm{C}}(x_1,\ldots,x_{k+3},r_n t) \text{ is connected}} \prod_{i=1}^{k+3} f(x_i)\mathrm{d}Vol_d(x_i) \\
&= \binom{n}{k+3}\int_M f(x_1)\mathrm{d}Vol_d(x_1)\int_{M^{k+2}} \mathbb{1}_{\check{\mathrm{C}}(x_1,\ldots,x_{k+3},r_n t) \text{ is connected}} \\
&\qquad \times \mathbb{1}_{x_2,\ldots x_{k+3}\in\mathcal{B}(x_1,(k+3)r_n t)^{k+2}} \prod_{i=2}^{k+3} f(x_i)\mathrm{d}Vol_d(x_i) \\
&\leq C_d\|f\|_\infty^{k+2}\binom{n}{k+3}\int_M f(x_1)((k+3)r_n t^+)^{d(k+2)}\mathrm{d}Vol_d(x_1) \\
&\leq C_{d,k,t^+,\|f\|_\infty}\binom{n}{k+3} r_n^{d(k+2)},
\end{aligned}$$

according to (Aamari and Levrard, 2019, Lemma B.7), since $(k+3)r_n t^+ \leq \tau_{\min}/4$ for n large enough. Thus,

$$\binom{n}{k+2}^{-1} r_n^{-d(k+1)}\mathbb{E}(L_{r_n,t}) \leq C_{d,k,t^+,\|f\|_\infty} n r_n^d.$$

Upper-bound of the variance

Let us denote by

$$U_n = \frac{1}{\binom{n}{m}}\sum_{I\subset\mathbb{X}_n,|I|=m} g_{s,t}(I),$$

where $m = k+2$, and, for $j = 1,\ldots,m-1$,

$$g_j(x_1,\ldots x_j) = \mathbb{E}\left(g_{s,t}(x_1,\ldots,x_j,X_{j+1},\ldots,X_m)\right).$$

We remark that $g_m = g_{s,t}$. Noting that U_n is a U-statistics of order m, Hoeffding's decomposition (see, e.g., (Lee, 1990, Theorem 3)) yields that

$$\mathrm{Var}(U_n) = \binom{n}{m}^{-1}\sum_{j=1}^{m}\binom{m}{j}\binom{n-m}{m-j}\mathrm{Var}(g_j). \tag{4}$$

Proceeding as for the bound on $\mathbb{E}(L_{r_n t})$, we may write

$$|g_j(x_1,\ldots x_j)| \leq C_{d,m,t^+,\|f\|_\infty} r_n^{d(m-j)}\mathbb{1}_{\check{\mathrm{C}}(x_1,\ldots x_j)\text{is connected'}}$$

so that

$$\mathrm{Var}(g_j(X_1,\dots,X_j)) \le \mathbb{E}(g_j^2(X_1,\dots,X_j) \le C_{d,m,t^+,\|f\|_\infty} r_n^{2d(m-j)+(j-1)d},$$

for $j = 1,\dots,m-1$. As well,

$$\mathrm{Var}(g_m(X_1,\dots,X_m)) \le \mathbb{E}(g_m^2(X_1,\dots,X_m)) \le C_{d,m,t^+,\|f\|_\infty} r_n^{(m-1)d}.$$

Plugging these inequalities into (4) leads to

$$\begin{aligned}
\mathrm{Var}(U_n) &\le \binom{n}{m}^{-1} C_{d,m,t^+,\|f\|_\infty} \sum_{j=1}^{m} \binom{m}{j}\binom{n-m}{m-j} r_n^{2d(m-j)+(j-1)d} \\
&\le C_{d,m,t^+,\|f\|_\infty} r_n^{d(2m-1)} \sum_{j=1}^{m} \frac{\binom{n-m}{m-j}}{\binom{n}{m}} r_n^{-dj} \\
&\le C_{d,m,t^+,\|f\|_\infty} r_n^{d(2m-1)} \sum_{j=1}^{m} \frac{1}{(nr_n^d)^j} \\
&\le C_{d,m,t^+,\|f\|_\infty} n^{-m} r_n^{d(m-1)},
\end{aligned}$$

for n large enough so that $\binom{n-m}{m-j}/\binom{n}{m} \le 2^j n^{-j}$ and $nr_n^d \le 1$. We deduce that

$$\begin{aligned}
\mathrm{Var}\left(\frac{G_n(s,t)}{r_n^{d(m-1)}\binom{n}{m}}\right) &= \frac{1}{r_n^{2d(m-1)}}\mathrm{Var}(U_n) \le C_{d,m,t^+,\|f\|_\infty}(n^m r_n^{d(m-1)})^{-1} \\
&\le C_{d,k,t^+,\|f\|_\infty}(n^{k+2} r_n^{d(k+1)})^{-1}.
\end{aligned}$$

Bounding the variance of $L_{r_n t}$ proceeds from the same calculation, noting that $L_{r_n t}$ is a U-statistic of order $m = k+3$. Namely, proceeding as above leads to

$$\mathrm{Var}\left(\frac{L_{r_n t}}{\binom{n}{m}}\right) \le C_{d,m,t^+,\|f\|_\infty} n^{-m} r_n^{d(m-1)},$$

with $m = k+3$, so that

$$\begin{aligned}
\mathrm{Var}\left(\frac{L_{r_n t}}{(r_n^d)^{k+1}\binom{n}{k+2}}\right) &\le C_{d,k,t^+,\|f\|_\infty} \frac{\binom{n}{k+3}^2}{\binom{n}{k+2}^2 r_n^{d(2k+2)}} n^{-(k+3)} r_n^{d(k+2)} \\
&\le C_{d,k,t^+,\|f\|_\infty} \frac{nr_n^d}{n^{k+2} r_n^{d(k+1)}} \le C_{d,k,t^+,\|f\|_\infty}(n^{k+2} r_n^{d(k+1)})^{-1},
\end{aligned}$$

for n large enough.

End of the proof

Let $k \le d-4$ and choose $r_n = n^{-\frac{k+2}{2+d(k+1)}}$. It holds

$$\begin{aligned}
r_n^2 &= n^{-(k+2)} r_n^{-d(k+1)} = n^{-\frac{2(k+2)}{2+d(k+1)}}, \\
nr_n^d &= n^{\frac{2+d(k+1)-d(k+2)}{2+d(k+1)}} = n^{\frac{2-d}{2+d(k+1)}} \le r_n.
\end{aligned}$$

The above calculation then leads to, for any $0 < s \leq t \leq u \leq v \leq t^+$, and n large enough,

$$\begin{aligned}\mathbb{E}\left[(\xi_{k,n} - \mu_k)(R_{s,t,u,v})^2\right] &\leq \mathbb{E}\left[(\xi_{k,n} - \mu_k)(R_{s,t,u,v})\right]^2 + \mathrm{Var}\left[(\xi_{k,n} - \mu_k)(R_{s,t,u,v})\right] \\ &\leq C_{d,k,t^+,\|f\|_\infty} n^{-(k+2)} r_n^{-d(k+1)} + C_{d,k,t^+,\|f\|_\infty} (L \vee \frac{\|f\|_\infty}{\tau_{\min}})^2 r_n^2 \\ &\leq C_{k,d,t^+,\|f\|_\infty,\tau_{\min},L} n^{-\frac{2(k+2)}{2+d(k+1)}}.\end{aligned}$$

Now, for $k \geq d-4$ and $r_n = n^{-\frac{k+4}{d(k+3)}}$, we get

$$\begin{aligned}n^2 r_n^{2d} = n^{-(k+2)} r_n^{-d(k+1)} = n^{-\frac{2}{k+3}}, \\ r_n = n^{-\frac{k+4}{d(k+3)}} \leq n^{-\frac{1}{k+3}} = n r_n^d.\end{aligned}$$

This yields, for n large enough,

$$\mathbb{E}\left[(\xi_{k,n} - \mu_k)(R_{s,t,u,v})^2\right] \leq C_{k,d,t^+,\|f\|_\infty,\tau_{\min},L} n^{-\frac{2}{k+3}}.$$

5.8 Proof of Corollary 11

Assume without loss of generality that the points are sampled according to f_1. Let $0 \leq s \leq t \leq u \leq v \leq \infty$. For $i = 1, 2$, denote by

$$l_i = \frac{\int_{\mathcal{M}} f_i^{k+2} \mathrm{d}Vol_d}{(k+2)!} \int_{(\mathbb{R}^d)^{k+1}} H_{s,t,u,v}(0, y_1, \ldots, y_{k+1}) dy_1 \ldots dy_{k+1}.$$

By Chebyshev's inequality,

$$\mathbb{P}\left(|\xi_{k,n}(R_{s,t,u,v}) - l_1| \geq \frac{|l_1 - l_2|}{2}\right) \leq \frac{4\mathbb{E}\left[(\xi_{k,n}(R_{s,t,u,v}) - l_1)^2\right]}{|l_1 - l_2|^2}.$$

Inverting the above formula and using the variance bound in the proof of Theorem 10 yields that there exists a constant C such that with probability greater or equal than $1 - C\frac{n^{-\frac{2(k+2)}{d(k+1)}}}{(\int_M |f_1^{k+2} - f_2^{k+2}|)^2}$,

$$|\xi_{k,n} - l_1| \geq \frac{|l_1 - l_2|}{2}.$$

This means that with at least the same probability, the data are correctly labeled as being sampled according to f_1.

References

Eddie Aamari and Clément Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *Ann. Statist.*, 47(1):177–204, 2019.

Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18, 2017.

Nima Aghaeepour, Greg Finak, Holger Hoos, Tim R Mosmann, Ryan Brinkman, Raphael Gottardo, and Richard H Scheuermann. Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*, 10(3):228–238, 2013.

Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial intelligence*, 201:81–105, 2013.

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The uea multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.

Francisco J Baldán and José M Benítez. Multivariate times series classification through an interpretable representation. *Information Sciences*, 569:596–614, 2021.

Christophe AN Biscio, Nicolas Chenavier, Christian Hirsch, and Anne Marie Svane. Testing goodness of fit for point processes via topological data analysis. *Electronic Journal of Statistics*, 14(1):1024–1074, 2020.

Omer Bobrowski and Robert J Adler. Distance functions, critical points, and the topology of random\v {C} ech complexes. *arXiv preprint arXiv:1107.4775*, 2011.

Omer Bobrowski and Matthew Kahle. Topology of random geometric complexes: a survey. *Journal of applied and Computational Topology*, 1(3):331–364, 2018.

Omer Bobrowski and Sayan Mukherjee. The topology of probability distributions on manifolds. *Probability theory and related fields*, 161(3):651–686, 2015.

Jean-Daniel Boissonnat, Frédéric Chazal, and Mariette Yvinec. *Geometric and topological inference*, volume 57. Cambridge University Press, 2018.

Peter Bubenik, Michael Hull, Dhruv Patel, and Benjamin Whittle. Persistent homology detects curvature. *Inverse Problems*, 36(2):025008, 2020.

Peter Bubenik et al. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16(1):77–102, 2015.

Mathieu Carrière and Andrew Blumberg. Multiparameter persistence image for topological machine learning. *Advances in Neural Information Processing Systems*, 33:22432–22444, 2020.

Mathieu Carriere, Marco Cuturi, and Steve Oudot. Sliced wasserstein kernel for persistence diagrams. In *International conference on machine learning*, pages 664–673. PMLR, 2017.

Mathieu Carrière, Frédéric Chazal, Yuichi Ike, Théo Lacombe, Martin Royer, and Yuhei Umeda. Perslay: A neural network layer for persistence diagrams and new graph topological signatures. In *International Conference on Artificial Intelligence and Statistics*, pages 2786–2796. PMLR, 2020.

Frédéric Chazal, Vin De Silva, and Steve Oudot. Persistence stability for geometric complexes. *Geometriae Dedicata*, 173(1):193–214, 2014a.

Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Convergence rates for persistence diagram estimation in topological data analysis. In *International Conference on Machine Learning*, pages 163–171. PMLR, 2014b.

Frédéric Chazal, Clément Levrard, and Martin Royer. Clustering of measures via mean measure quantization. *Electronic Journal of Statistics*, 15(1):2060–2104, 2021.

Mark De Deuge, Alastair Quadros, Calvin Hung, and Bertrand Douillard. Unsupervised feature learning for classification of outdoor 3d scans. In *Australasian conference on robitics and automation*, volume 2, page 1. University of New South Wales Kensington, Australia, 2013.

Persi Diaconis, Susan Holmes, Mehrdad Shahshahani, et al. Sampling from a manifold. *Advances in modern statistical theory and applications: a Festschrift in honor of Morris L. Eaton*, 10:102–125, 2013.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

Herbert Edelsbrunner and John L Harer. *Computational topology: an introduction*. American Mathematical Society, 2022.

Herbert Federer. Curvature measures. *Trans. Amer. Math. Soc.*, 93:418–491, 1959. ISSN 0002-9947. doi: 10.2307/1993504. URL `https://doi.org/10.2307/1993504`.

Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*, volume 76. Springer, 2006.

Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.

Felix Hensel, Michael Moor, and Bastian Rieck. A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, 4:681108, 2021.

Matthew Kahle. Random geometric complexes. *Discrete & Computational Geometry*, 45 (3):553–573, 2011.

Varun Khurana, Harish Kannan, Alexander Cloninger, and Caroline Moosmüller. Supervised learning of sheared distributions using linearized optimal transport. *Sampling Theory, Signal Processing, and Data Analysis*, 21(1):1–51, 2023.

Johannes TN Krebs and Wolfgang Polonik. On the asymptotic normality of persistent betti numbers. *arXiv preprint arXiv:1903.03280*, 2019.

Genki Kusano, Yasuaki Hiraoka, and Kenji Fukumizu. Persistence weighted gaussian kernel for topological data analysis. In *International conference on machine learning*, pages 2004–2013. PMLR, 2016.

Tam Le and Makoto Yamada. Persistence fisher kernel: A riemannian manifold kernel for persistence diagrams. *Advances in Neural Information Processing Systems*, 31, 2018.

A. J. Lee. *U-statistics*, volume 110 of *Statistics: Textbooks and Monographs*. Marcel Dekker, Inc., New York, 1990. ISBN 0-8247-8253-4. Theory and practice.

Clément Maria, Jean-Daniel Boissonnat, Marc Glisse, and Mariette Yvinec. The gudhi library: Simplicial complexes and persistent homology. In *International congress on mathematical software*, pages 167–174. Springer, 2014.

Andreas Maurer. A chain rule for the expected suprema of gaussian processes. *Theoretical Computer Science*, 650:109–122, 2016.

Shahar Mendelson and Roman Vershynin. Entropy and the combinatorial dimension. *Inventiones mathematicae*, 152(1):37–55, 2003.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Caroline Moosmüller and Alexander Cloninger. Linear optimal transport embedding: Provable wasserstein classification for certain rigid transformations and perturbations. *arXiv preprint arXiv:2008.09165*, 2020.

Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. *Advances in neural information processing systems*, 25, 2012.

Ippei Obayashi, Yasuaki Hiraoka, and Masao Kimura. Persistence diagrams with linear machine learning models. *Journal of Applied and Computational Topology*, 1(3):421–449, 2018.

Takashi Owada. Convergence of persistence diagram in the sparse regime. *arXiv e-prints*, art. arXiv:2103.12943, March 2021.

Takashi Owada. Convergence of persistence diagram in the sparse regime. *arXiv preprint arXiv:2103.12943*, 2021.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Barnabás Póczos, Aarti Singh, Alessandro Rinaldo, and Larry Wasserman. Distribution-free distribution regression. In *Artificial Intelligence and Statistics*, pages 507–515. PMLR, 2013.

Raphael Reinauer, Matteo Caorsi, and Nicolas Berkouk. Persformer: A transformer architecture for topological machine learning. *arXiv preprint arXiv:2112.15210*, 2021.

Jan Reininghaus, Stefan Huber, Ulrich Bauer, and Roland Kwitt. A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4741–4748, 2015.

Martin Royer, Frédéric Chazal, Clément Levrard, Yuhei Umeda, and Yuichi Ike. Atol: measure vectorization for automatic topologically-oriented learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1000–1008. PMLR, 2021.

Sivan Sabato and Naftali Tishby. Multi-instance learning with any hypothesis class. *The Journal of Machine Learning Research*, 13(1):2999–3039, 2012.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1): 5272–5311, 2016.

Michel Talagrand. Vapnik–chervonenkis type conditions and uniform donsker classes of functions. *The Annals of Probability*, 31(3):1565–1582, 2003.

Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3460–3469, 2015.

D Yogeshwaran, Eliran Subag, and Robert J Adler. Random geometric complexes in the thermodynamic regime. *Probability Theory and Related Fields*, 167(1):107–142, 2017.