From topological features to machine learning models : a journey through persistence diagrams

Olympio Hacquard

université PARIS-SACLAY

From topological features to machine learning models : a journey through persistence diagrams

De la topologie aux méthodes d'apprentissage automatique : utiliser puis dépasser les diagrammes de persistance

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 574, Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées
Graduate School : Mathématiques. Référent : Faculté des sciences d'Orsay

Thèse préparée au **Laboratoire de Mathématiques d'Orsay** (Université Paris-Saclay, CNRS), sous la direction de **Gilles BLANCHARD**, professeur, et le co-encadrement de **Clément LEVRARD**, professeur.

Thèse soutenue à Paris-Saclay, le 15 septembre 2023, par

Olympio HACQUARD

Composition du jury

Membres du jury avec voix délibérative

Elisabeth GASSIAT Professeure, Université Paris-Saclay	Présidente
Marc HOFFMANN Professeur, Université Paris-Dauphine	Rapporteur
Ulrike VON LUXBURG Professeure, University of Tübingen	Rapporteure
Yasuaki HIRAOKA Professeur, Kyoto University	Examinateur
Bertrand MICHEL Professeur, Ecole Centrale Nantes	Examinateur

THESE DE DOCTORAT
NNT : 2023UPASM019

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre: De la topologie aux méthodes d'apprentissage automatique : utiliser puis dépasser les diagrammes de persistance

Mots clés: Analyse topologique des données, diagrammes de persistance, théorie de l'apprentissage statistique, régression, classification

Résumé: La raison d'être de l'analyse topologique des données est d'extraire de l'information de nature topologique afin d'aider à analyser des jeux de données. Cette information peut alors être aisément incluse dans une chaîne de traitement pour effectuer diverses tâches d'apprentissages sur les données. Un des objets les plus présents dans ce cadre est le diagramme de persistance. Mathématiquement, cet objet est une mesure discrète où les coordonnées de chaque point correspondent à des échelles auxquelles une composante topologique est présente dans les données. Supposons que l'on ait accès à des observations bruitées d'une fonction lisse, le diagramme de persistance peut alors être scindé en une composante de bruit et une composante de signal. La première contribution de cette thèse est d'exploiter cette information pour un problème de régression afin de reconstruire une fonction bruitée. En minimisant un critère topologique, on parvient à annuler le bruit et récupérer un signal lisse. Cependant, cette dichotomie entre le signal et le bruit est assez grossière, et les diagrammes de persistance contiennent beaucoup d'information pouvant permettre de classifier des données. En raison de leur structure de mesures, ils ne peuvent être mis tels quels en entrée d'algorithmes d'apprentissage automatique standards. La deuxième contribution de cette thèse est de proposer une méthode de classification de mesures, et l'adaptation des principes fondateurs de la théorie de l'apprentissage statistique dans ce contexte. On a également contribué à l'étude asymptotique des diagrammes de persistance dans un cadre aléatoire. En pratique, l'information utile contenue dans les diagrammes peut être redondante et on peut être intéressés par seulement quelques statistiques bien choisies extraites du diagramme. Dans une troisième contribution, on a développé des descripteurs basés sur le calcul de la caractéristique d'Euler, qui sont bien plus rapides à calculer que les diagrammes de persistance, tout en conservant une performance similaire. De plus, ces descripteurs peuvent être adaptés à une évolution multi-paramètre de la topologie des données, permettant de dépasser une restriction théorique des diagrammes de persistance qui contraint à l'utilisation d'un seul paramètre d'évolution.

université PARIS-SACLAY

ÉCOLE DOCTORALE
de mathématiques
Hadamard (EDMH)

Title: From Topological Features to Machine Learning Models: A Journey through Persistence Diagrams

Keywords: Topological data analysis, persistence diagrams, statistical learning theory, regression, classification

Abstract: Topological data analysis consists in extracting topological information to help analyze data sets. This information can be easily included in a workflow to perform various learning tasks on the data. One of the most ubiquitous objects is the persistence diagram. It is a discrete measure where the coordinates of each point correspond to a range of scales at which a topological feature is present in the data. Assuming we observe a noisy observation of a smooth function, persistence diagrams can be separated into a noise and a signal component. The first contribution of this thesis is to use this information in a regression framework to estimate a noisy function. By minimizing a topological criterion, we manage to cancel the noise and retrieve a smooth signal. However, such a noise-signal dichotomy is very coarse, and persistence diagrams contain much information that can help classify data. As they are measures, they cannot be input as such in a standard machine learning pipeline. In a second contribution, we have developed a method that classifies measure data and adapted elements of statistical learning theory in this framework. We have also contributed to the study of the asymptotic persistence diagrams built in a random setting. In practice, we are only interested in using limited information from persistence diagrams. In a final contribution we demonstrate that a few well chosen statistics are enough to obtain competitive accuracy in classification problems. These descriptors are typically based on the Euler characteristic and are much faster to compute than persistence diagrams since we do not compute the coordinates of all the points in the diagram. Furthermore, these descriptors can be adapted to a multi-valued evolution of the topology of the data, going beyond a theoretical limitation of persistence diagrams that restricts to the use of a single evolution parameter.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France


université
PARIS-SACLAY
FACULTÉ
DES SCIENCES
D'ORSAY


MATHEMATIQUES
ORSAY


Fondation mathématique
FMJH
Jacques Hadamard


Inria

Remerciements

En finalisant le présent ouvrage, j'ai pu faire la rétrospective de ce qu'il s'est passé ces trois dernières années, extrêmement riches et fécondes à bien des égards. J'ai notamment pu réaliser à quel point rien de tout cela n'aurait pu être possible sans les gens qui m'entourent. Ce manuscrit, c'est aussi un peu le vôtre !

Mes premiers remerciements vont à mes deux encadrants Gilles et Clément pour leur patience, leur implication et leur dévouement sans faille, bravant confinements et décalages horaires pour me proposer un encadrement 5 étoiles ! Vous avez réussi à trouver un parfait équilibre dans lequel je me suis senti accompagné et aidé tout en conservant une totale liberté dans mes directions de recherche. Je souhaite également remercier Fred qui a joué un véritable rôle d'impresario dès la genèse de cette thèse, en me faisant découvrir l'analyse topologique de données, puis en m'aidant à trouver un stage puis une thèse. Speaking of which, I am extremely grateful to Wolfgang and Krishna for their great implication and the work we have accomplished together. Thanks again for the invitation in Davis, I will for sure remember these few months overseas!

I would also like to thank the referees Ulrike and Marc, as well as the jury: Elisabeth, Bertrand and Yasu for taking the time to review this thesis.

J'ai également une pensée pour toutes les belles rencontres que j'ai pu faire au LMO, à Datashape et à Davis, merci les collègues ! En particulier merci à JB, mon alcolyte de conférences et à Vadim avec qui on en aura passé des journées à faire de la cuisine. Ces petits rendez-vous hebdomadaires ont changé pour le meilleur la vision que j'avais de la recherche.

Merci également à toutes les personnes qui m'ont ouvert leurs portes pendant ces trois années d'errance, d'Arroyo Grande à Gisenyi en passant par les quatre coins de France. Si jamais ces quelques lignes vous parviennent, sachez que vous avez joué un rôle dans ma construction, en m'apprenant ce qu'on ne peut pas trouver dans les bouquins de maths.

Un énorme merci à la bande de gros singes du Manoir : Jean-Jean, Cre, Pif, Cox, Glier, Alice, Bastos, Rickton, Beubeul, RBK, Colin et à tous ceux qui gravitent autour, amis de plus ou moins longue date. On a partagé beaucoup plus qu'une maison, et les souvenirs que j'y garde vont me marquer à vie.

Quoi de plus logique que de finir par ceux qui me sont le plus cher : Papa, Maman, Daph, Val, Ambroise, Tarek et les trois petits monstres Casper, Sabin et Virgile. Merci pour votre amour !

Contents

I Introduction (English)

The explosion of data availability of all natures has been the source of a genuine revolution. Whether generating texts, classifying images or forecasting time series, data sciences have triggered a keen interest in developing corresponding mathematical tools. One of the current challenges in machine learning is the study of high-dimensional data sets, especially when a small number of data is available. Defying all intuition as illustrated in the first chapter of [Gir14], these data sets often make classical methods fail. However, some data are often considered to lie in a much simpler structure. In particular, the *manifold assumption* states that many high-dimensional data sets from real-world applications are supported on low-dimensional non-flat structures. The most stereotypical example is that of natural images: apparently having a dimension equal to the number of pixels, the numerous constraints (large constant zones, edges, corners) firmly lower the number of degrees of freedom such that images from a specific data set are often considered to live in a low-dimension manifold.

Developing tools to analyze such geometric data sets has been the motto of topological data analysis (TDA). Using techniques from algebraic topology, TDA extracts topological and geometric information from all sorts of data sets. The objective of applied TDA is twofold:

- Building topological descriptors from the data in order to achieve an automated machine learning task such as classification, regression, or clustering, see [CM21].
- Help bringing a qualitative understanding of the data from a topological perspective, see [Hes20] and [RB19].

Finding its origins in the works on persistent homology of [ELZ00] and [CZCG04], topological data analysis has encountered a sought-after success due to its wide variety of applications, for instance, in health, [RYB+20, FM22, ACC+21], neurology [KDS+18], biology, [IOH20, RB19], material sciences, [LBD+17, HNH+16], cosmology [PEVdW+17] and more recently music theory, [AAPL22, MBP22].

One of the most ubiquitous objects in TDA is the *persistence diagram* which summarizes all the topological information contained in the data. This descriptor will be the unifying thread of this dissertation and we will study how to use it to perform various data analysis tasks and how to overcome its limitations. This dissertation is based on the three following research papers:

- The article [HBB+22], joint with Gilles Blanchard, Krishnakumar Balasubramanian, Clément Levrard and Wolfgang Polonik, where we study how to use persistence diagrams to enforce smoothness in a regression set-up.
- The preprint [HBL23], joint with Gilles Blanchard and Clément Levrard, where we study structural properties of persistence diagrams, and develop new tools in measure classification to extract information from the diagrams and classify data.
- The preprint [HL23], joint with Vadim Lebovici, where we study descriptors that achieve a better performance than persistence diagrams at a lower computational cost, while preserving some interpretability properties.

After a short introduction to topological data analysis and its key concepts in Section I.1, we will expose the ideas and main contributions of each of these papers in Sections I.2, I.3 and I.4. Each main section of this dissertation then corresponds to a different article.

I.1 A short introduction to topological data analysis and persistence diagrams

We refer to the textbooks [EH22] and [BCY18] for a detailed introduction to computational topology and topological data analysis, and we recall the principal concepts and notions, as well as some typical strategies in TDA.

I.1.1 Simplicial homology and persistence diagrams

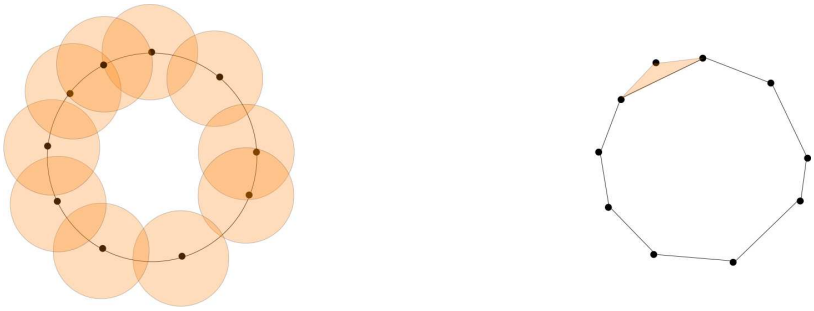
Before introducing the foundations of persistence theory, we start with a few notions on simplicial complexes in order to build intuition about the topological nature of data in a simple discrete framework:

Definition I.1. A *(finite) abstract simplicial complex* $\mathcal{K}$, or *simplicial complex*, is a finite collection of finite sets that is closed under taking subsets. An element $\sigma \in \mathcal{K}$ is called a *simplex*, and subsets of σ are called *faces* of σ. The dimension of a simplicial complex is the maximal dimension of one of its simplices.

It can be shown, see Chapter III of [EH22], that by mapping each abstract vertex to a point in $\mathbb{R}^{2d+1}$, every simplicial complex of dimension d has a geometric realization in $\mathbb{R}^{2d+1}$, where a *(geometric) k-dimensional simplex* is the convex hull of $k+1$ affinely independent points. A 0 (resp. 1, 2, 3)-dimensional simplex is called a vertex (resp. an edge, a triangle, a tetrahedron). One of the primary examples of geometric simplicial complex built over a point cloud is the Čech complex:

Definition I.2. Let $\mathbb{X} \subseteq \mathbb{R}^d$ be finite. The *Čech complex at scale* $t \geq 0$ is the simplicial complex $\check{\mathcal{C}}(\mathbb{X}, t)$ defined as follows: for $(x_0, \ldots, x_k) \in \mathbb{X}^{k+1}$, the simplex $\{x_0, \ldots, x_k\}$ is in $\check{\mathcal{C}}(\mathbb{X}, t)$ if the intersection of closed balls $\cap_{l=0}^{k} \overline{B}(x_l, t)$ is non-empty.

In Figure 1, taken from [Wik23], we illustrate the construction of the Čech complex for a small sample of points on a circle.



(a) Balls centered on each point (b) Corresponding Čech complex

Figure 1: Construction of a Čech complex for points sampled on a circle.

We see on Figure 1b that this simplicial complex looks like a "simplified" circle as it is connected and has a single large cycle.

Questions I.3. The two natural questions that come to mind are the following:

- How to algebraically characterize a cycle and the fact that the simplicial complex here is connected and only has a single cycle?

- Here, the union of balls has the same topology as a cycle for a range of scales t, including the one from Figure 1a. However, if the scale taken is typically larger than the circle's radius or smaller than the largest distance between two consecutive points, the topology of the union of balls will be different from that of a circle. How to choose the radius of the balls, i.e. the scale parameter of the Čech complex, without any prior knowledge of the data?

To answer the first question, we must introduce some notions of *simplicial homology*. We start with the definitions of a k-chain and of the boundary operator.

Definition I.4. Let $\mathcal{K}$ be a simplicial complex. A *simplicial k-chain* is a finite formal sum $\sum_{i=1}^{N} c_i \sigma_i$ where each σ_i is a k-simplex from $\mathcal{K}$ and $c_i \in \mathbb{Z}/2\mathbb{Z}$. The group of $k-$chains is denoted by C_k.

Definition I.5. The *boundary operator* $\partial_k : C_k \to C_{k-1}$ is a homomorphism defined by:

$$\partial_k(\sigma) = \sum_{i=0}^{k} \{v_0, \dots, \hat{v_i}, \dots, v_k\},$$

where $\sigma = \{v_0, \dots, v_k\}$, and the $(k-1)$-dimensional simplex $\{v_0, \dots, \hat{v_i}, \dots, v_k\}$ is the face of σ obtained by removing the vertex v_i. We further define two subgroups of C_k: the group of cycles $Z_k = \text{Ker } \partial_k$ and the group of boundaries $B_k = \text{Im } \partial_{k+1}$.

For instance, the boundary of a triangle is the sum of its three edges. It is an easy exercise to check that $\partial_{k+1}\partial_k = 0$, i.e. that the boundary of the boundary of a chain is always equal to 0. This implies that B_k is a subgroup of Z_k. Intuitively, a topological "k-dimensional hole" is a cycle of dimension k, which is not the boundary of a simplicial subcomplex. Topological holes can be rigorously defined in the following way:

Definition I.6. The *k-th homology group* H_k of $\mathcal{K}$ is defined as the quotient abelian group $H_k(\mathcal{K}) = Z_k/B_k$. Its rank $\beta_k = \text{rank } (H_k(\mathcal{K}))$ is called the *k-th Betti number*.

Intuitively, the k-th Betti number β_k counts the number of k-dimensional holes for $k \geq 1$, and β_0 is the number of connected components. Back to the example of Figure 1b, there is a long chain that is not a boundary; thus $\beta_1 = 1$. Since the simplicial complex is connected here, we have $\beta_0 = 1$. The Betti numbers of a d-dimensional simplex appear as signatures of the topology of a simplicial complex. In that vein, another invariant is given by the *Euler characteristic*:

Definition I.7. The *Euler characteristic* of a d-dimensional simplicial complex $\mathcal{K}$ is the integer defined as

$$\chi(\mathcal{K}) = \sum_{\sigma \in \mathcal{K}} (-1)^{\dim \sigma} = \sum_{k=0}^{d} (-1)^k \beta_k(\mathcal{K}).$$

The Euler characteristic is less informative than the collection of all β_k for $k \in \{0, \dots, d\}$. However, its expression as a simplices count makes it a much simpler invariant than the Betti numbers and is, therefore, much faster to compute in practice. Still, we will extensively use the Euler characteristic in Section V and demonstrate that it is a very powerful descriptor in data analysis when considered with a multi-scale lens for a family of simplicial complexes.

We can now carry on to the second question in I.3, which is the purpose of *persistent homology*. A high-level answer to this question would be that the best way of picking a radius parameter for the balls of the Čech complex would be to consider all possible radii. Formally, we start by defining a *filtration* of simplicial complexes:

Definition I.8. Consider a finite simplicial complex $\mathcal{K}$ and a non-decreasing function $f : \mathcal{K} \to \mathbb{R}$, in the sense that $f(\sigma) \leq f(\tau)$ whenever σ is a face of τ. We have that for every $a \in \mathbb{R}$, the sublevel set $\mathcal{K}(a) = f^{-1}((-\infty, a])$ is a simplicial subcomplex of $\mathcal{K}$. Considering all possible values of f leaves us with a nested family of subcomplexes

$$\varnothing = \mathcal{K}_0 \subset \mathcal{K}_1 \subset \ldots \subset \mathcal{K}_n = \mathcal{K},$$

called a *filtration*, where $a_0 = -\infty < a_1 < a_2 < \ldots < a_n$ are the values taken by f on the simplices of $\mathcal{K}$.

The Čech complex previously introduced on a point set $\mathbb{X}$ naturally defines a filtration over the complete simplicial complex $\mathcal{K} = 2^{\mathbb{X}}$:

$$\varnothing \subset \check{\mathcal{C}}(\mathbb{X}, 0) = \mathbb{X} \subset \check{\mathcal{C}}(\mathbb{X}, t_1) \subset \check{\mathcal{C}}(\mathbb{X}, t_2) \subset \ldots \subset \check{\mathcal{C}}(\mathbb{X}, t_n) = \mathcal{K},$$

where $0 < t_1 < t_2 < \ldots < t_n$ are the times at which we observe a change of topology for the Čech complex. This filtration is simply called the *Čech filtration* and denoted by $\check{\mathcal{C}}(\mathbb{X})$. The simplicial complex $\mathcal{K}$ being finite, the number of critical times t_i for the Čech complex is finite. Furthermore, it can be shown, see [BE17], that these times are radii of circumscribed balls of simplices of $\mathcal{K}$.

Here, the idea is to consider a nested family of simplicial complexes and to track when topological features get created and destroyed, i.e. the evolution of Betti numbers. This can be formalized in the following way: for $i \leq j$, the canonical inclusion map $\mathcal{K}_i \to \mathcal{K}_j$ induces a homomorphism $f_k^{i,j} : H_k(\mathcal{K}_i) \to H_k(\mathcal{K}_j)$ on the homology groups for every k, therefore leading to a sequence of homology groups:

$$0 = H_k(\mathcal{K}_0) \to H_k(\mathcal{K}_1) \to \ldots \to H_k(\mathcal{K}_n) = H_k(\mathcal{K}).$$

We then define the *persistent homology*:

Definition I.9. The k-th *persistent homology groups* are the images of the homomorphisms defined above: $H_k^{i,j} = \text{im } f_k^{i,j}$. Similarly to homology, we define the $k-$th *persistent Betti number* by $\beta_k^{i,j} = \text{rank } H_k^{i,j}$.

As the name suggests, the k-th persistent Betti number $\beta_k^{i,j}$ corresponds to the number of k-dimensional holes that *persist* between $\mathcal{K}_i$ and $\mathcal{K}_j$. Given a class γ in $H_k(\mathcal{K}_i)$, we say that this class if *born* at $\mathcal{K}_i$ if $\gamma \notin H_k^{i-1,i}$. Similarly, the death time of γ (possibly infinite) is the smallest index j such that $f_k^{i,j}(\gamma) \in H_k^{i-1,j}$. We are now in a position to define the object that is at the core of this dissertation:

Definition I.10. The k-th *persistence diagram* is the multi-set of $\overline{\mathbb{R}}^2$ of (birth, death) coordinates for every class γ that exists in the persistent homology sequence.

This multi-set can be turned into a discrete measure, introducing a convenient viewpoint in Section IV. Furthermore, note that the persistent Betti number $\beta_k^{i,j}$ equals the number of points in the infinite upper-left quadrant with angle (a_i, a_j). In Figure 2, adapted from [RCL+21], we can see an example of the construction of the persistence diagram of a Čech complex built over a point cloud.

Here, the points are arranged near two circles, and we can see two cycles. The topology of the point cloud itself is constituted of as many connected components as the number of points in the input cloud. In order to construct the Čech complex of this point cloud, we

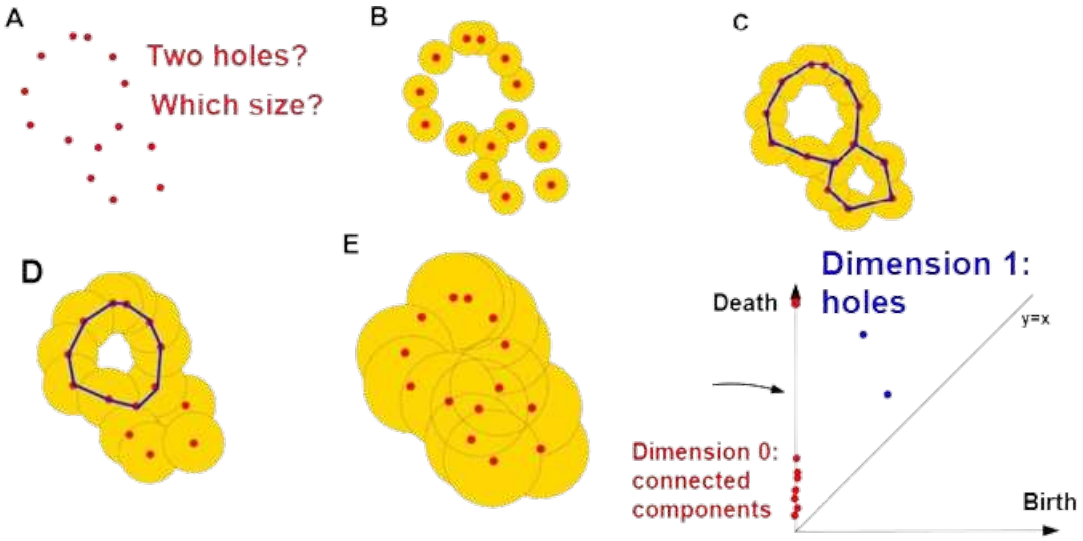


Figure 2: Persistence diagram of the Čech complex of a point cloud

center balls around each point and let the radius grow from 0 to ∞. One after the other, balls of points next to each other will start having a non-empty intersection, creating edges in the Čech filtration and thus merging connected components. Topologically, a merging corresponds to the death of one of the components. Thus, the persistence diagram for the 0-homology (connected components) contains points of coordinates $(0, d_i)$ and a point with coordinates $(0, \infty)$ since when the Čech radius is large enough (Figure 2, E), the union of balls remains connected. As for the cycles, the union of balls does not contain any cycle for smaller radii (Figure 2, B). As the radius becomes large enough (Figure 2, C), we see two cycles appearing in the Čech complex. The critical radius at which a cycle starts existing is its *birth time*. Finally, when the radius grows as in Figure 2, D and E, three-fold intersections of balls create triangles in the Čech complex that fill the cycles. The corresponding critical radius is the *death time* of the cycle. Finally, this leaves us with two points in the persistence diagram for the 1-homology. In addition to characterizing the topology of a point cloud (here that it has two cycles), persistence diagrams also give information about the geometric size of the holes. Furthermore, information about the sampling and the inter-distance between points can be extracted from the death times of the 0-dimensional topological features.

The success of topological data analysis comes from the possibility of comparing persistence diagrams. The most popular way is to compute the *bottleneck distance* defined as:

Definition I.11. Let $\Delta = \{(x, x) | x \in \mathbb{R}\}$ be the diagonal of $\mathbb{R}^2$.

The *bottleneck distance* d_B between two persistence diagrams D and D' is defined by:

$$d_B(D, D') = \inf_{\eta : D \cup \Delta \to D' \cup \Delta} \sup_{x \in D \cup \Delta} \|x - \eta(x)\|_\infty,$$

where the infimum is taken over all bijections η from $D \cup \Delta$ to $D' \cup \Delta$.

This distance inspired by optimal transport can be generalized to any p-Wasserstein distance. We illustrate the optimal matching between two diagrams in Figure 3, adapted from [Cha23]. The bottleneck distance here is the length of the longest edge (in infinity norm). We

refer to [DL21] for more information about the structure of the space of persistence diagrams endowed with such metrics. A strength of persistence diagrams is their stability property, see [CSEH07], which states that a small perturbation in the input data induces a small perturbation of the persistence diagrams in terms of bottleneck distance. We will state the precise result in a specific case in Section I.1.2.

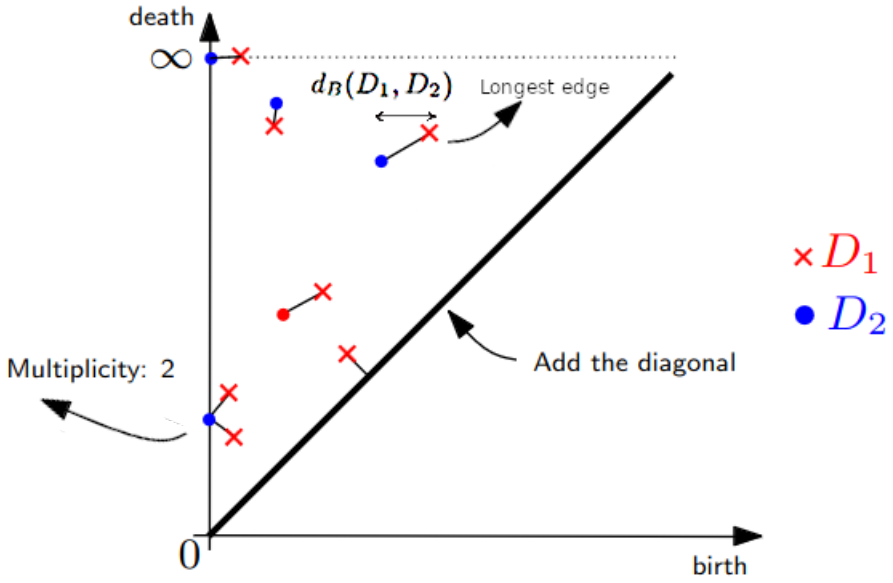


Figure 3: Optimal matching of two persistence diagrams with different numbers of points

Thanks to its multi-scale approach, persistence diagrams carry much information of a topological nature about the input data. These descriptors can then be used to perform learning tasks such as classification, regression, or clustering. We refer to the survey [HMR21] for applying persistence diagrams to machine learning in practice. Because of its representation as a multi-set of $\overline{\mathbb{R}}^2$ (or a discrete measure), persistence diagrams are not suited for classical machine learning algorithms. A typical strategy is to turn these descriptors into features in a Banach space, as done with Betti curves [RSL20], persistence images [AEK+17], landscapes [B+15], and more recently measure-oriented vectorizations in [RCL+21] and neural network methods in [CCI+20, RCB21]. We will see in Section IV a method to perform statistical learning directly in the space of measures without any vectorization.

On the numerical side, simplicial complexes, persistent homology, and persistence diagrams are computed using the `Gudhi` library, [MBGY14]. The first algorithm to compute persistence is in [ZC04]. Computing a persistence diagram for a simplicial complex $\mathcal{K}$ has worst-case time complexity $\mathcal{O}(|\mathcal{K}|^\omega)$ where $2 \leq \omega < 2.373$ is the exponent for matrix multiplication; see [MMS11]. Although computing persistence diagrams is reasonable, having an algorithm that scales well for large complexes or in high dimensions remains an open question. Indeed, for a typical complex, the number of simplices grows exponentially with the dimension.

Finally, persistence diagrams can be computed for a much larger class of input than Čech complexes of point clouds. In Section V, we will explore several filtrations on point clouds, while in Sections IV and V, we will see how to apply it to graph data. This framework can also be used on images and 3D volumes by computing *cubical complexes* instead of simplicial

complexes. This approach has been booming to the study of medical images, for instance in [ACC+21] or [JKN20]. Furthermore, the theory of persistent homology extends beyond finite complexes. It can be applied to more general nested sequences of topological spaces, always to track the birth and death times of topological features. The leading example is the study of sublevel sets $f^{-1}((-\infty, t])$ of a Morse function $f : \mathbb{R}^d \to \mathbb{R}$ that is indeed a nested sequence of subspaces of $\mathbb{R}^d$. This example motivates the Section III of the present work, and we will expose some of its aspects in Section I.1.2.

I.1.2 Topological persistence for Morse functions

We start this section by defining Morse functions and will demonstrate why they are particularly suitable in topological data analysis.

Definition I.12. Let $f : \Omega \to \mathbb{R}$ be a $\mathcal{C}^2$ function where Ω is a subset of $\mathbb{R}^d$.

- A *critical point* of f is a point $x \in \Omega$ such that $\nabla f(x) = 0$. The corresponding value $f(x)$ is called a *critical value*.
- Let x_0 be a critical point of f. Its *index* is the number of negative eigenvalues of the Hessian matrix of f at x_0.

Definition I.13. We say that a real-valued function f defined on a differentiable manifold is *Morse* if it is $\mathcal{C}^2$ and all its critical points are non-degenerate, in the sense that the Hessian matrix at each critical point is non-degenerate.

Note that the set of Morse functions is an open dense subset of smooth functions. One of the main properties of Morse functions is summed up in the following theorem, adapted from Theorems 3.1 and 3.2 from [Mil63]:

Theorem I.14. *Let f be a Morse function on a smooth manifold $\mathcal{M}$ and denote by $\mathcal{M}^a$ the sublevel set $f^{-1}((-\infty, a])$.*

- *Assume there is no critical value between $a < b$. Then $\mathcal{M}^a$ and $\mathcal{M}^b$ are diffeomorphic and $\mathcal{M}^b$ deformation retracts onto $\mathcal{M}^a$.*
- *Assume p is a non-degenerate critical point of f with index s and that $f(p) = q$. We further assume there are no other critical points p' with $f(p') = q$. Then for ε small enough, $\mathcal{M}^{q+\varepsilon}$ is homotopy equivalent to $\mathcal{M}^{q-\varepsilon}$ with a s-handle attached.*

This theorem illustrates that topological changes in the sublevel sets occur at critical values for Morse functions. Similarly to simplicial homology, let us now consider the evolution of the homology groups of sublevel sets $f^{-1}((-\infty, t])$ as t traverses $\mathbb{R}$ from $-\infty$ to $+\infty$, and note in a persistence diagram the values at which topological components are born and die. We illustrate this in Figure 4, adapted from [CM21], where we consider the height function of a "nosy torus" surface with its 0 and 1-persistence diagrams, respectively in red and blue.

Consider a plane of equation $z = t$ and let t grow from $-\infty$ to $+\infty$. The sublevel sets are all empty before reaching the global minimum a_0. Then in a_0, a connected component is born, and the sublevel sets are cup-shaped. Another connected component is created at the local minimum a_1, and the two connected components eventually merge at the saddle point a_3. By a convention called the *Elder rule*, we consider that the youngest component dies first, hence a point at coordinate (a_1, a_3) in the 0-dimensional persistence diagram. Starting from a_2, the sublevel sets contain a 1-cycle, and in a_4, a second cycle gets created, corresponding

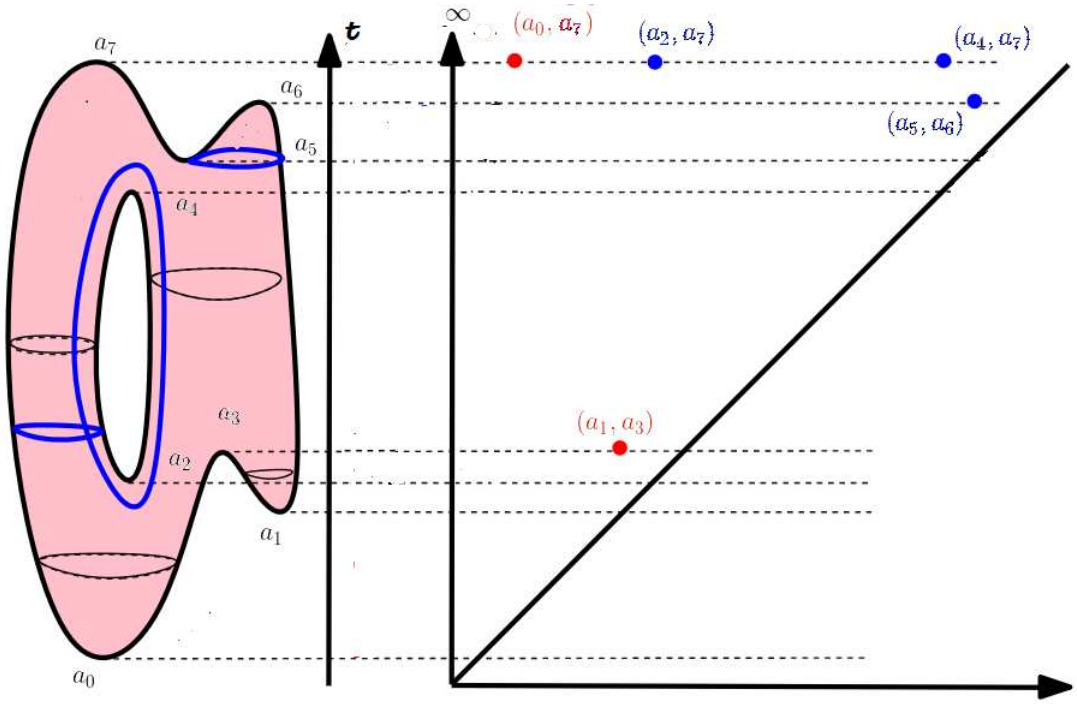


Figure 4: Sublevel sets filtration of a nosy torus and corresponding persistence diagrams

to the homological features of a torus. A third cycle is created at the saddle-point a_5, which only persists until the local maximum a_6 closes it, hence a point of coordinates (a_5, a_6) in the 1-persistence diagram. After a_7, the sublevel sets have the same topology as the torus. They remain connected forever so that the connected component initially created in a_0 never dies. Similarly, the two cycles created in a_2 and a_7 never dies. As a convention in order to avoir point with infinite coordinates, we set their death times to the maximum of the function, a_7.

Here, we see the impact of Theorem I.14 for computing persistence diagrams since the coordinates of all the points in the persistence diagram are critical values of the height function. More precisely, for this 2-dimensional function, minima always create connected components while maxima always kill cycles. As for saddle points, they can either merge two connected components or give birth to a cycle.

We are now armed with a multi-set that describes a Morse function and its critical points from a topological perspective. Being interested in applications in statistics, a natural question that comes to mind is the behaviour of this descriptor when adding noise to the function. If we add a noise bounded by ε, we can expect each critical point to be displaced from at most ε, and many critical points that do not persist more than ε are created. We make this fact more precise in the following *stability result*, taken from [CSEH07], in terms of the bottleneck distance introduced in Definition I.11. We denote by D_f the persistence diagram for the sublevel sets of a Morse function f.

Theorem I.15. *Let $\mathcal{M}$ be a topological space and let f and g be two Morse functions from $\mathcal{M}$ to $\mathbb{R}$. We have that:*

$$d_B(D_f, D_g) \leq \|f - g\|_\infty.$$

Having such a stable metric over the space of diagrams seems very interesting at first glance for data analysis because it shows some robustness to bounded noise. In our context, we would investigate how to *denoise* a function using such topological regularizers. To do so, we introduce the *persistence* of a function:

Definition I.16.
- For a feature (b, d) of a persistence diagram, its *persistence* is equal to its lifetime $d - b$.
- The *persistence* of a diagram D is equal to the sum of all individual lifetimes:

$$\mathrm{Pers}(D) = \sum_{(b,d)\in D} (d - b).$$

When there is no possible confusion, we denote by $\mathrm{Pers}_k(f)$ the persistence of the k-th persistence diagram of the sublevel sets of a given function f, and by $\mathrm{Pers}(f) = \sum_k \mathrm{Pers}_k(f)$ its *total persistence*.

We illustrate in Figure 5 the behaviour of persistence and bottleneck distances when adding Gaussian noise to measurements of a smooth function. In this setup, we consider observations $Y_i = f^\star(X_i) + \varepsilon_i$ where points $(X_i)_{i=1}^{1000}$ are uniformly sampled on the unit square, $f^\star$ is the function displayed in Figure 5a which is a sum of four Gaussians, and $\varepsilon \sim \mathcal{N}(0, \sigma I_n)$. We display an interpolation of the observations in Figures 5a and 5c. When adding noise to each measurement, many critical points with a small lifetime are created and mapped to the diagonal when computing the bottleneck distance, which illustrates the stability Theorem I.15 with high probability. The bottleneck distance here measures the noise added to the function. However, it requires knowledge of the true regression function $f^\star$. Similarly, 0-persistence and 1-persistence measure how much noise has been added to a function. There are stability results in some sense for persistence, but they involve the number of points in the persistence diagrams, which can only be roughly bounded. We will make this claim more precise later in Lemma III.2.

We now have all the tools to move on to the first contribution where we investigate the use of total persistence in a regression setting.

I.2 Enforcing regularity using persistence diagrams

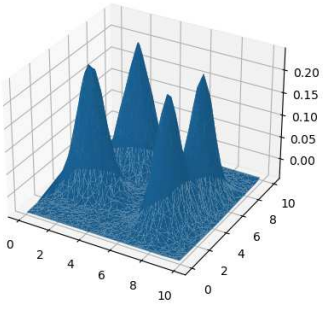
Total persistence captures small oscillations of a function and is therefore indicated as a regularizer in a regression framework. This observation is at the core of section III, published in [HBB+22]. The goal here will be to use total persistence as a penalty term, in order to cancel the observation noise, resulting in a smoother and more accurate prediction. This work focuses on a regression setting where we observe data on a manifold.

I.2.1 Topological regularization

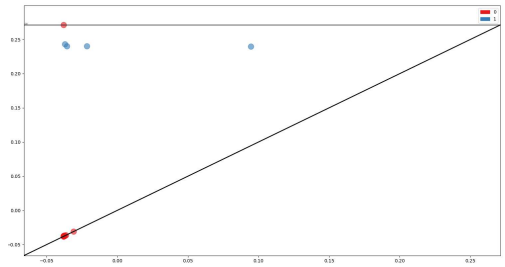
We consider a regression problem on a compact manifold $\mathcal{M}$. We assume data points $(X_i)_{i=1}^n$ are sampled uniformly and independently over $\mathcal{M}$ and that we observe real responses:

$$Y_i = f^\star(X_i) + \varepsilon_i,$$

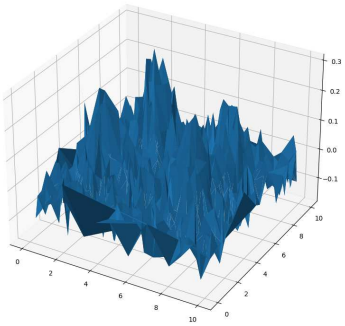
where $(\varepsilon_i)_{1\le i\le n}$ are i.i.d. zero-mean sub-Gaussian noise variables independent of all the X_i's. Our goal is to retrieve the function $f^\star$ to denoise the input or predict the response's value at an unobserved data point.



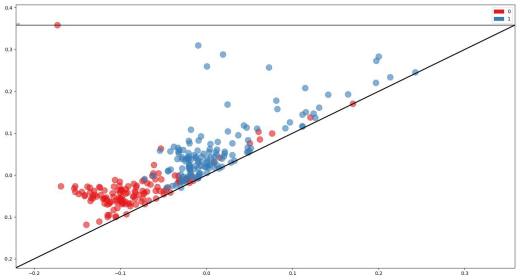
(a) Original function



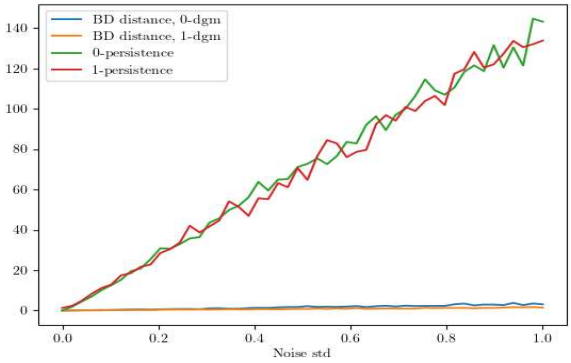
(b) Corresponding persistence diagram



(c) Noisy function



(d) Corresponding persistence diagram



(e) Bottleneck distances and persistences as functions of the noise standard deviation

Figure 5: Stability of bottleneck distance and persistence to Gaussian noise

We first start by considering a basis of functions $(\Phi_i)_{i\geq 0}$ adapted to the manifold, called the eigenfunctions of the *Laplace-Beltrami operator*, see [Ros97]. This basis is a generalization of the Fourier basis of functions to compact manifolds. In practice, we do not have access to these eigenfunctions $(\Phi_i)_{i\geq 0}$ in most cases. However, we can approximate them using the spectrum of the graph Laplacian matrix, see [Chu97], which can be computed easily from the data. In order to find the best coefficients for approximating $f^\star$ with the first p eigenfunctions of the Laplace-Beltrami operator (or graph Laplacian), we minimize the following criterion:

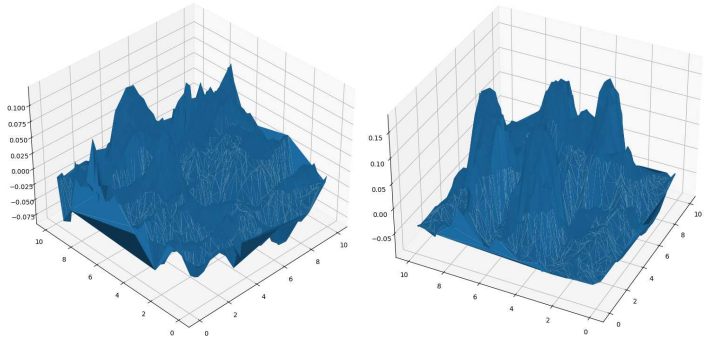
$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \theta_j \Phi_j(x_i) \right)^2 + \mu \Omega(\theta),$$

where Ω is a penalty term introduced to prevent overfitting and is calibrated by the scalar μ.

Examples of classical penalties include L^1-regularization, also called Lasso (see [Bv11] for an exhaustive treatment), L^2-regularization also called Ridge regression, [HK70], or total variation penalty, [HR16], which is the most similar to our work since it aims at penalizing large oscillations. In our work, we have considered two penalties:

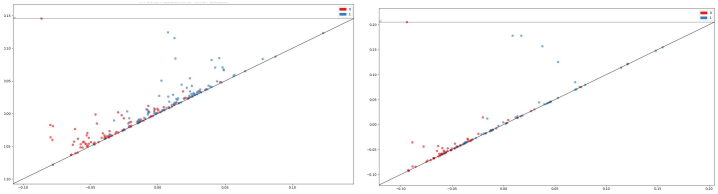
- $\Omega_1(\theta) = \sum_{j=1}^{p} |\theta_j| \text{Pers}(\Phi_j)$ is a weighted-lasso type penalty, fast and easy to implement, and which acts as a variable selector, favoring eigenfunctions with a small persistence, i.e. that do not oscillate too much.
- $\Omega_2(\theta) = \text{Pers}\left(\sum_{j=1}^{p} \theta_j \Phi_j\right)$ is a non-convex penalty where we used techniques from [CCG+21] to minimize the loss function $\mathcal{L}$. Following the logic previously described in Figure 5, this penalty acts as a denoizer, removing low-persistence points from the persistence diagram, thus providing a smoother estimate of the function.

In practice, we combine these two penalties by selecting a subset of eigenfunctions thanks to Ω_1 and then perform denoising using the regularizer Ω_2. We can see the effect of the Ω_2 regularizer itself on Figure 6 where we denoise the function from Figure 5c. The reconstruction from a noisy sample with a topological penalty is compared to a reconstruction on the Laplace eigenbasis with a simple Lasso penalty. We can see that the reconstruction penalized by Ω_2 is smoother, and although there has been a loss of information, we manage to reconstruct the four peaks of the original signal. This smoothing is attested by the persistence diagrams: the topological reconstruction has four very persistent features and very little topological noise, as opposition to the Lasso reconstruction, which has only two significantly persistent features, and a fair number of points close to the diagonal, resulting in a much coarser estimate.



(a) Function estimated by a Lasso

(b) Function estimated by the topological penalty Ω_2



(c) Persistence diagram of the Lasso estimate

(d) Persistence diagram of the Ω_2 penalty estimate

Figure 6: Reconstruction of the sum of four Gaussians.

I.2.2 Detailed list of contributions

The contributions of this work are the following:

(i) We perform a thorough qualitative and quantitative analysis on synthetic and real data, comparing topologically penalized methods to other standard regression methods. Our method is comparable to Kernel Ridge Regression, [Vov13] and outperforms other regularization techniques, including total variation.

(ii) We provide several oracle results for the penalties Ω_1 and Ω_2. In particular, we show that assuming that the regression function writes as $f^\star = \sum_{j=1}^{p} \theta_j^\star \Phi_j$, the optimal parameter $\theta^\star$ is approximated at a rate $O(p/n)$. In addition, we have a theoretical result on the persistence of the reconstructed function that guarantees its "topological smoothness".

(iii) We provide a negative result about the capacity of the set of functions of bounded persistence.

This work is the first occurrence of theoretical guarantees for topologically regularized models. Although the experimental results are promising, they have a major computational drawback. Indeed, in order to minimize $\mathcal{L}$ with penalty Ω_2, we need to compute hundreds of persistence diagrams (namely one per gradient step) which turns out to be quite costly. In addition, we have adopted a very simplistic point of view in the analysis of persistence diagrams, that we split between a signal component assumed to correspond to the diagram of the true function $f^\star$, and a clump of points near the diagonal that we assume correspond to the noise ε and that we aim at destroying by minimizing the total persistence. We will now see that much information about the data is contained close to the diagonal and can be used for classification purposes.

I.3 Reading information in the low-persistence features

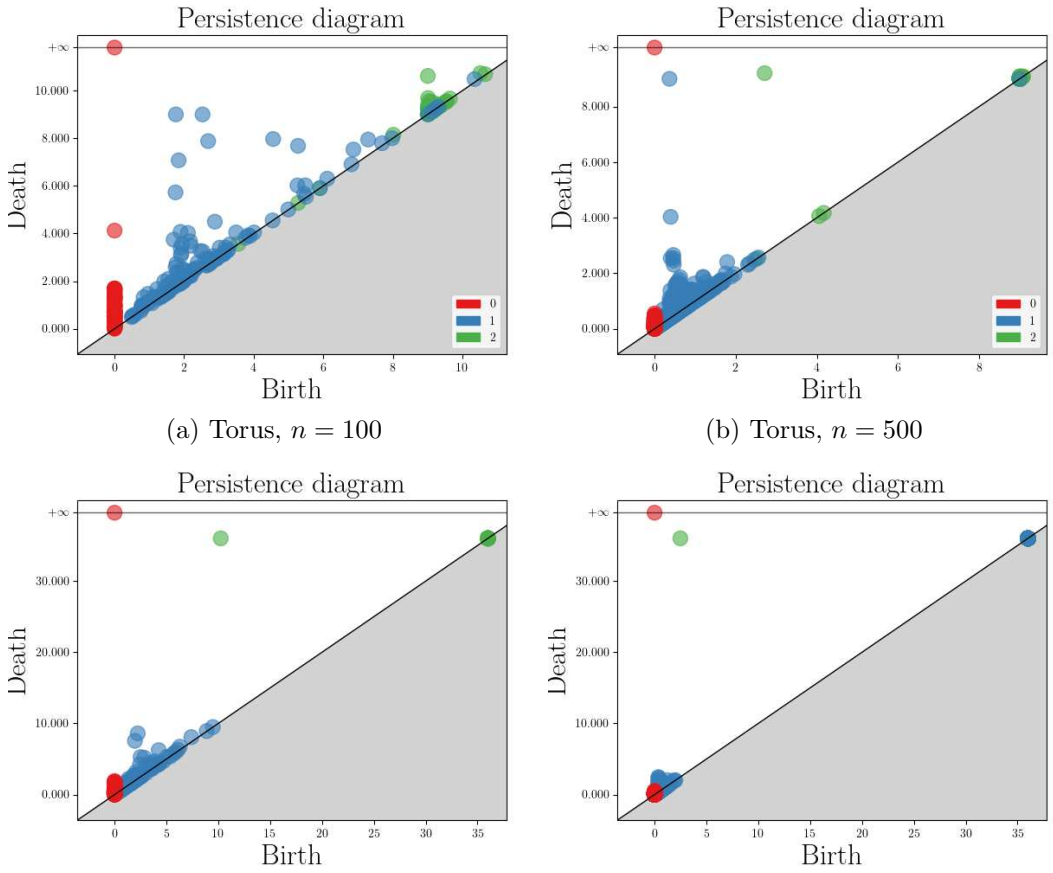
We now focus on the case of persistence diagrams built with the Čech filtration of a point cloud in $\mathbb{R}^d$ and how to classify them. This work has led to the preprint [HBL23]. We consider a supervised binary classification problem where we observe data $D_N = (\mu_1, Y_1), \dots, (\mu_N, Y_N)$ where μ_i is a persistence diagram and Y_i is a label in $\{0, 1\}$. Typically, this originates from a classification problem on raw data turned into persistence diagrams (e.g. graphs, images, time series or point clouds) because it is assumed that they carry some topological information that can be relevant for discriminating between two classes.

I.3.1 The torus versus sphere flagship example

As a toy example, consider the problem of discriminating between point clouds on a sphere $\mathbb{S}^2$ and point clouds on a torus $\mathbb{T}^2$. One way to tackle this problem in a translation and rotation invariant way is to construct the 1-persistence diagrams of the Čech filtrations of the data. We have that $\beta_1(\mathbb{T}^2) = 2$ and $\beta_1(\mathbb{S}^2) = 0$. Stated otherwise, the torus has two independent cycles while any closed loop on the sphere retracts onto a point. In Figure 7, we show an example of persistence diagrams of the Čech filtration over a point cloud of varying size sampled uniformly on a torus or a sphere.

Therefore, it is to be expected that as long as we sample enough points, the 1-persistence diagram of the Čech filtration of the point cloud on a torus has two very persistent cycles, while that of a sphere has no cycle that persists for a long time.

Due to their structure as a multi-set of points (or a discrete measure on $\overline{\mathbb{R}}^2$), persistence diagrams are not suited for standard classification algorithms. As previously discussed, a



(a) Torus, $n = 100$ (b) Torus, $n = 500$

(c) Sphere, $n = 100$ (d) Sphere, $n = 500$

Figure 7: 0, 1 and 2-persistence diagrams for n points uniformly sampled.

common strategy is to map persistence diagrams in a Banach space, [DP19] and turn them into vectors of a given size. We propose to solve the two following problems:

- Is it possible to circumvent this vectorization step and directly develop a classification method that takes measures as input while having good theoretical guarantees?
- In this torus versus sphere experiment, likewise to Section I.2, we again treat the persistence diagrams as high-persistence features corresponding to homological components of the underlying manifold and low-persistence features that correspond to noise. Can we go beyond this topological signal-noise dichotomy and what relevant information lies in the low-persistence features?

I.3.2 The measure-classification problem

We develop in Section IV a vectorization-free method to classify measures in a supervised fashion. Roughly speaking, for measures defined on some compact metric space $\mathcal{X}$, we look at different zones of $\mathcal{X}$ and discriminate according to the mass put in each zone. For instance, in the torus versus sphere problem, looking at whether the discrete point measure associated with the persistence diagram puts more than two points above a certain death level is enough to classify it as a torus. We provide two algorithms to learn these zones and the corresponding activation thresholds, which are then aggregated using a *boosting* procedure. In addition, we

describe a broader class of measure classifiers and derive statistical guarantees. Roughly speaking, consider a class of functions $\mathcal{F}$ on $\mathcal{X}$, and a corresponding class $\tilde{\mathcal{F}}$ defined on the space $\mathcal{M}(\mathcal{X})$ of measures on $\mathcal{X}$ by

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

We relate capacity measures of $\tilde{\mathcal{F}}$ in terms of corresponding quantities for $\mathcal{F}$, which are, in general, much simpler to compute. More precisely, for a class of function $\mathcal{F}$, we define the empirical Rademacher complexity on a sample $(Z_i)_{i=1}^N$ as

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

where $(\sigma_1, \ldots, \sigma_N)$ are independent Rademacher random variables, i.e. for every i, $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$. This quantity measures the correlation of the class of function $\mathcal{F}$ with a vector of random Rademacher noise. Under some mild assumptions, the predictive risk is usually upper-bounded by the Rademacher complexity with high probability; see [SSBD14]. Having a low Rademacher complexity, therefore, translates into good generalization properties. In our case, we have established the following upper-bound:

Theorem I.17. *There exists an absolute constant K such that*

$$\mathcal{R}_N(\tilde{\mathcal{F}}) \leq \frac{K\overline{M}_2\sqrt{VC(\mathcal{F})}}{\sqrt{N}},$$

where $\overline{M}_2 = \left(\frac{1}{N}\sum_{i=1}^N \mu_i(\mathcal{X})^2\right)^{\frac{1}{2}}$, and VC is another capacity measure called the Vapnik-Chervonenkis dimension. We denote that the empirical Rademacher complexity of $\tilde{\mathcal{F}}$ is upper-bounded by a capacity measure on the class $\mathcal{F}$, which is usually more informative and much simpler to compute. In Section IV, we will establish a lower-bound of the same order and see how these results can translate into comprehensive prediction bounds for the classifiers on the space of measures we have built.

I.3.3 Asymptotic results for Čech complexes of random point clouds

Back to persistence diagrams, it turns out that in many practical situations, our algorithm will find that the most discriminatory zones lie near the diagonal, which means that some relevant information is contained in what we unjustly called the *topological noise*. Some solid theoretical guarantees back this observation up. Indeed, in [BHPW20], the authors claim that information about the curvature of the underlying space can be extracted from low-persistence features. Furthermore, these features can be analyzed to extract information about the sampling density.

Given n points $\mathbb{X}_n = (X_1, \ldots, X_n)$ i.i.d. sampled according to a density f on $\mathbb{R}^d$, it is a very natural question to wonder about the convergence of the persistence diagrams of the Čech complex of $\mathbb{X}_n$ as n tends to infinity. We refer to [BK18] for a survey of existing limit results for topological quantities. We must introduce a re-scaling sequence $(r_n)_{n\in\mathbb{N}}$ that tends to 0 to make the limit non-trivial. The speed at which (r_n) tends to 0 is crucial: denote by $\Lambda := \lim_{n\to\infty} nr_n^d \in [0, \infty]$. In order to illustrate our claim, we consider the most favourable case $\Lambda = 0$ called the *sparse regime*, and cite a result from [Owa22]:

Theorem I.18. *Let $\mathbb{X}_n = (X_1, \dots, X_n)$ be an i.i.d. sample drawn according to an a.e. continuous bounded Lipschitz density g on $\mathbb{R}^d$. Consider a sequence $(r_n)_{n \in \mathbb{N}}$ such that we are in the sparse regime $nr_n^d \to 0$. Further assume that $n^{k+2} r_n^{d(k+1)} \to \infty$ as $n \to \infty$. Denote by $\xi_{k,n}$ the k-th persistence diagram of $\check{\mathcal{C}}(\frac{1}{r_n}\mathbb{X}_n)$. 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*

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

for $0 < s \leq t \leq u \leq v$, where H is a geometric function that depends on the mutual positions of its arguments. Then, we have the vague convergence:

$$\frac{\xi_{k,n}}{n^{k+2} r_n^{d(k+1)}} \xrightarrow{v} \mu_k \text{ almost surely.}$$

Similar results are cited in [Owa22] for other sub-regimes of the sparse regime $n^{k+2} r_n^{d(k+1)} \to 0$ and $n^{k+2} r_n^{d(k+1)} \to c > 0$.

Remark I.19. We can make a few comments about this rather surprising result.

- The limiting measure μ_k only depends on the sampling density through the multiplicative constant $\int_{\mathbb{R}^d} f^{k+2}$, showing some universality of the limiting object.
- This constant $\int_{\mathbb{R}^d} f^{k+2}$ is a global quantity, although persistence diagrams account for local effects.
- Assume we try to classify whether a point cloud has been generated according to a density f_1 or a density f_2 such that there exists k such that $\int_{\mathbb{R}^d} f_1^{k+2} \neq \int_{\mathbb{R}^d} f_2^{k+2}$. Counting the number of points of the persistence diagrams that fall in any rectangle (after a suited rescaling) will identify the correct model when the number of sample points is large enough. In particular, for this classification task where there is no homological signal to retrieve, the "topological noise" contains discriminative information about the sampling.

I.3.4 Detailed list of contributions

The contributions of this work are the following:

(i) We have proposed a vectorization-free method to classify measures, have implemented it and tested it against benchmark methods on topological data analysis data sets, and also on time series and flow cytometry data sets.
(ii) We have developed a theory encompassing this method and given lower and upper bounds for the Rademacher complexity of the class of functions $\tilde{\mathcal{F}}$ previously defined.
(iii) We have derived theoretical guarantees, specifically in the case of persistence diagrams classification, to discriminate either between manifolds with different homology or samplings on the same manifold. This work has been the occasion to generalize Theorem I.18 to samplings on manifolds while using an alternative proof technique.

I.4 Beyond persistence diagrams: Euler tools and multi-persistence

We have now demonstrated that there is relevant information both away and close to the diagonal. Another drawback of the method of Section I.2 is its computational cost. Indeed,

to minimize the total persistence Ω_2, we need to compute a persistence diagram per gradient step. In addition, for a persistence diagram D, we were only interested in the quantity $\sum_{(b,d)\in \mathrm{D}}(d-b)$, so that computing the entire diagram and the exact coordinates of every point clearly overshoots the problem. It does not seem that there is a way to circumvent this technical issue. However, this observation has led to the following question:

Question I.20. What type of statistics on persistence diagrams can be computed in linear time from the filtration values? How can they be used in a data analysis context?

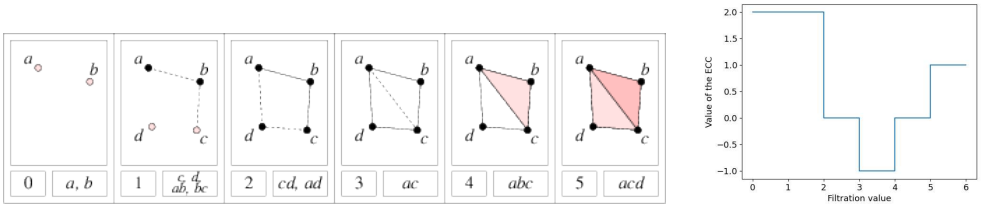
I.4.1 Euler characteristic curves and their integral transforms

Question I.20 is the object of Section V and has been pre-published in [HL23]. To find descriptors that do not require computing persistence, let us return to the Euler characteristic of a simplicial complex defined in Definition I.7. We have noticed that for a simplicial complex $\mathcal{K}$ of dimension d, computing $\sum_{k=0}^{d}(-1)^k\beta_k(\mathcal{K})$ can be done very simply by counting simplices thanks to the formula of Definition I.7. On the other hand, there is no such trick to compute $\sum_{k=0}^{d}\beta_k(\mathcal{K})$ and each β_k needs to be computed individually using the homology groups of $\mathcal{K}$, which is significantly more costly. Therefore, alternating over homological dimensions seems key to accessing fast descriptors. In that logic, we define the *Euler characteristic curve (ECC)*:

Definition I.21. Consider a finite simplicial complex $\mathcal{K}$ and a filtration $(\mathcal{K}_t)_{t\in\mathbb{R}}$. The *Euler characteristic curve* is the function

$$\chi_{\mathcal{K}} : t \in \mathbb{R} \mapsto \chi(\mathcal{K}_t) \in \mathbb{Z}.$$

In Figure 8, we show the construction of the Euler characteristic curve for a simple filtration taken from [ZC04].



(a) Filtration (b) Corresponding ECC

Figure 8: Construction of the Euler characteristic curve of a filtration.

In practice, the Euler characteristic curve is a finite vector used as a feature to perform various learning tasks such as classification, regression, or clustering. Although this seems like a very simple descriptor, it has often been used in applications, see [SZ21, JKN20, AQO+22], and we will demonstrate in Section V that it has a strong predictive power comparable to that of persistence diagrams for a much reduced computational cost.

In order to extract relevant information from the Euler characteristic curves, we consider integral transforms of these functions. This object has been studied theoretically in [Leb22] under the denomination of *hybrid transforms (HT)*.

Definition I.22. Let $(\mathcal{K}_t)_{t\in\mathbb{R}}$ be a filtration of a simplicial complex $\mathcal{K}$ and $\chi_{\mathcal{K}}$ its corresponding Euler curve. Let $\kappa \in L^1(\mathbb{R})$. The *hybrid transform* with kernel κ is the function defined by:

$$\psi^\kappa : \xi \in \mathbb{R}_+^\star \mapsto \xi \cdot \int_{\mathbb{R}} \kappa(\xi s)\chi_{\mathcal{K}}(s)\,\mathrm{d}s.$$

Up to multiplication by ξ, hybrid transforms are classical integral transforms of Euler curves. For instance, if $\kappa = \cos$, it corresponds to the cosine Fourier transform. We will demonstrate that hybrid transforms constitute robust descriptors in data analysis once discretized, especially for unsupervised problems. In addition, the following lemma connects them directly to persistence diagrams and provides a partial answer to Question I.20:

Lemma I.23. *Let $(\mathcal{K}_t)_{t\in\mathbb{R}}$ be a filtration, and ψ^κ its hybrid transform with kernel κ. Let $\overline{\kappa}$ be a primitive of κ such that $\overline{\kappa}(x) \underset{x\to\infty}{\longrightarrow} 0$. Denote by $D_k = \{(b_i^k, d_i^k)\}_{i=1,\dots,n_k}$ the k-th persistence diagram of $(\mathcal{K}_t)$. For a simplex $\sigma \in (\mathcal{K}_t)$, we denote by $t(\sigma)$ the first time it appears in the filtration. We therefore have, for every $\xi \in \mathbb{R}_+^\star$:*

$$\psi^\kappa(\xi) = \sum_{k\geq 0}\sum_{i=1}^{n_k}(-1)^k\Big(\overline{\kappa}(\xi\cdot b_i^k) - \overline{\kappa}(\xi\cdot a_i^k)\Big). \tag{I.1}$$

In addition, we have:

$$\psi^\kappa(\xi) = -\sum_{\sigma\in\mathcal{K}}(-1)^{\dim\sigma}\overline{\kappa}\big(\xi\cdot t(\sigma)\big). \tag{I.2}$$

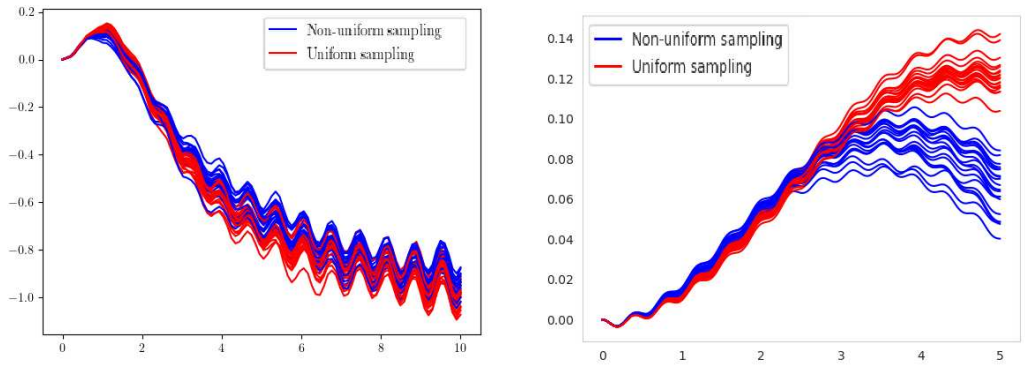
This lemma shows that for every $\overline{\kappa}$ that vanishes at infinity, the function defined by Equation (I.1) can be computed linearly in terms of filtration values by Equation (I.2). If $\overline{\kappa}$ does not vanish at infinity, we can still use Equation (I.2) as a proxy for the hybrid transform with kernel κ. Finally, in addition to bringing fast and powerful classification features without computing any diagram, hybrid transforms also bring a gain in understanding the structure of the topological noise. Indeed, in Section I.3, we have demonstrated that low-persistence features for the Čech complex of a point cloud sampled on a manifold carry information about the sampling density and local quantities of the manifold. However, how they manifest explicitly in the topological noise is still unclear. We can use hybrid transforms to go beyond this interpretation in the following classification problem.

Consider 500 points sampled on a torus embedded in $\mathbb{R}^3$. The first class corresponds to points uniformly sampled on the torus; see [DHS+13]. The second class corresponds to drawing two angles (θ, φ) uniformly in $[0, 2\pi]^2$ and obtain a point on the torus through the embedding $\Psi_{\mathbb{T}^2} : (\theta, \varphi) \mapsto (x_1, x_2, x_3)$, where:

$$\begin{cases} x_1 = (2 + \cos(\theta))\cos(\varphi), \\ x_2 = (2 + \cos(\theta))\sin(\varphi), \\ x_3 = \sin(\theta). \end{cases}$$

Note that this does not produce a uniform sampling on the torus. We consider a similar set-up on the sphere, where one class corresponds to 500 points sampled uniformly. At the same time, for the other, we draw 500 angles θ uniformly in $[0, \pi]$ and φ according to a normal distribution centred on π. We obtain a point on the sphere via the classical spherical coordinates parametrization $\Psi_{\mathbb{S}^2} : (\theta, \varphi) \mapsto (x_1, x_2, x_3)$ where:

$$\begin{cases} x_1 = \sin(\theta)\cos(\varphi), \\ x_2 = \sin(\theta)\sin(\varphi), \\ x_3 = \cos(\theta). \end{cases}$$



(a) HT, torus data

(b) HT, sphere data

Figure 9: ECC and HT, two sampling on a torus and sphere

In Figure 9, we plot the beginning of hybrid transforms with kernel $\kappa = \sin$ of the Čech complex of a few samples of each class for both set-ups.

We remark that the oscillations of the transforms are in phase and have the same amplitude. However, from one manifold to another, the phase and amplitude of the oscillations of the transforms differ significantly. It suggests that they are related to global quantities and are signatures of the support manifold. In contrast, the samplings show up in the vertical shifts of the oscillations of the transforms. This experiment is the first step towards a more thorough understanding of the quantities involved in the low-persistence features.

I.4.2 The challenge of multi-persistence

So far, we have only discussed filtrations on simplicial complexes where the filtration function is real-valued, see Definition I.8. However, we may want to study the sublevel sets of multiple functions defined on a simplicial complex. In topological data analysis, taking multi-parameter filtrations is known as *multi-persistence*. The most prominent example couples the Čech filtration repeatedly mentioned and a function on the point cloud itself, such as a density estimator, see [CB20] for instance. In this case, the density estimator can filter the outliers that would make the Čech filtration unstable. This simple and natural extension is the source of many theoretical and practical challenges in persistence theory. Most importantly, there is no equivalent to the persistence diagram for multi-persistence; see [CZ09]. There have been a few attempts to design vectorizations adapted to bi-persistence, such as persistence landscapes in [Vip20], and persistence images in [CB20]. Similarly, our tools naturally generalize to multi-persistence: the Euler characteristic curve becomes a *Euler characteristic profile*, see [DG22]. Even though there is no more equivalent to Equation (I.1), hybrid transforms can still be computed using an analogue of Equation (I.2), and they now differ from classical integral transforms. This method remains very competitive in complexity and allows us to use up to five different filtrations in some settings. Indeed, if we aim at classifying graph data, many natural functions can suggest structural differences between graphs from different classes. More than enhanced and faster vectorizations of persistence diagrams, Euler characteristic profiles and hybrid transforms, therefore, appear as a necessity to go beyond persistence diagrams as the latter imposes the use of a single filtration function.

I.4.3 Theoretical guarantees

Finally, we conclude this section with some theoretical guarantees for Euler characteristic profiles and hybrid transforms. These results are of two different natures, and we start by mentioning those related to the *stability* of these descriptors, in the particular case of sublevel set filtrations defined over the same simplicial complex:

Stability Assume that we filter a finite simplicial complex $\mathcal{K}$ with two functions $f, g : \mathcal{K} \to \mathbb{R}^m$ and denote by χ_f and χ_g the corresponding Euler profiles and by ψ_f^κ and ψ_g^κ the corresponding hybrid transforms of bounded integrable kernel κ. We have the following stability lemma for these two descriptors:

Lemma I.24. *Let $M > 0$. We have that*

$$\|(\chi_f - \chi_g)\mathbb{1}_{[-M,M]^m}\|_1 \leq (2M)^{m-1}\|f - g\|_1.$$

In addition, let $q \in [1, \infty]$. There exists a constant C depending only on q such that

$$\|\psi_f^\kappa - \psi_g^\kappa\|_q \leq C\|\kappa\|_\infty\|f - g\|_1.$$

In this lemma, the L^1 norm of a function f defined on a simplicial complex $\mathcal{K}$ is defined as $\|f\|_1 = \sum_{\sigma\in\mathcal{K}} \|f(\sigma)\|_1$. This lemma shows that these two descriptors demonstrate some robustness to perturbations of filtrations. However, this stability only holds for the L^1 norm, that involves the total number of simplices in the complex $\mathcal{K}$. This stability result is therefore weaker than the bottleneck stability stated in Theorem I.15 that holds for the L^∞ distance.

Limit theorems In addition to these stability results, we show that the descriptors from this section verify some guarantees in an asymptotic setting. In the case of Euler characteristic curves, this has been deeply studied in the literature, and we can cite [KRP21] for a functional central limit theorem for the Euler characteristic curve in a random setting. In the case of hybrid transforms, Equation I.1 states that for a given kernel κ and a mono-filtration $(\mathcal{K}_t)$ having persistence diagrams D_k in homological dimension $k \in \{0, \dots, d-1\}$, we have that

$$\psi^\kappa(\xi) = \sum_{k=0}^{d-1} \langle D_k, h_\xi \rangle,$$

where $h_\xi : (x, y) \mapsto \kappa(\xi y) - \kappa(\xi x)$. The persistence diagrams are seen as discrete measures as in Section I.3. This simple observation has the two following benefits for our work:

- The formulation of Section I.3.2 can be applied to hybrid transforms for a family of kernels.
- Known asymptotic results for persistence diagrams translate in asymptotic guarantees for hybrid transforms under some mild assumption on the kernel κ. More precisely, we can cite a result in the case of mono-persistence that combines Theorem I.18 with the above observation:

Theorem I.25. *Let $X_1, \dots, X_n$ be an i.i.d. sample drawn according to an a.e. continuous bounded Lipschitz density g on $\mathbb{R}^d$. Consider a sequence $(r_n)_{n\in\mathbb{N}}$ such that $nr_n^d \to 0$ and $n^{k+2}r_n^{d(k+1)} \to \infty$ as $n \to \infty$. We denote by ψ_n^κ the hybrid transform of the Čech filtration associated with the rescaled sample $\frac{1}{r_n}(X_i)_{i=1}^n$. Let $T, a > 0$ and $\kappa \in L^1(\mathbb{R})$. Further assume*

that κ is supported on $[0,T]$. Then there exist functions $A_0,\dots,A_{d-1}$ on $\mathbb{R}_+^*$ that depend only on κ such that for every $\xi > a$,

$$\frac{1}{n^{k+2} r_n^{d(k+1)}} \cdot \psi_n^{\kappa}(\xi) \underset{n\to\infty}{\longrightarrow} \sum_{k=0}^{d-1} \frac{(-1)^k}{(k+2)!} \cdot A_k(\xi) \cdot \int_{\mathbb{R}^d} g^{k+2}(x)\,\mathrm{d}x \qquad a.s..$$

As in Theorem I.18, the sampling density only appears through the global quantities $\int_{\mathbb{R}^d} g^{k+2}$ for $k = 0, 1, \dots, d-1$. Likewise to persistence diagrams, hybrid transforms can discriminate between different sampling densities as soon as n is large enough at a reduced computational cost. Finally, in Section V, we will establish similar limit results for multi-persistence also inherited from Equation (I.1).

I.4.4 Detailed list of contributions

Our contributions regarding this work are the following:

(i) We perform an in-depth qualitative study of Euler characteristic profiles and hybrid transforms in various settings while discussing the choice of the kernel parameter for hybrid transforms.

(ii) We demonstrate that Euler profiles achieve state-of-the-art accuracy in supervised classification and regression tasks when coupled with a robust classifier, such as a random forest, at a meagre computational cost.

(iii) We demonstrate that hybrid transforms act as highly efficient information compressors, similar to Fourier transforms. Consequently, they outperform Euler profiles in unsupervised classification tasks and supervised tasks when plugging a linear classifier. We also illustrate their ability to capture fine-grained information on a real-world data set.

(iv) We provide several theoretical guarantees for these descriptors. First, we prove stability properties in the flavour of Theorem I.15 that clarify the robustness of our tools concerning perturbations. Expressed in terms of L_1 norms, these are also hints of the sensitivity of our tools to the underlying geometry of the data at hand. Then, we establish the pointwise convergence of hybrid transforms associated with Čech filtrations of random samples and their asymptotic normality. We also establish a law of large numbers in a quite general multi-filtration set-up.

Outline

The dissertation is organised as follows: Section II is the translation of this introduction in French. In Section III, we study regression problems on manifolds using a topological penalty. This work is based on the topological signal-noise paradigm discussed in Section I.2. This is joint work with Krishnakumar Balasubramanian, Gilles Blanchard, Clément Levrard and Wolfgang Polonik, published in [HBB+22]. In Section IV, we develop a method to perform supervised classification on measure data. We provide two algorithms that fit within a more general framework of statistical learning on measures, for which we establish several theoretical guarantees. This work puts a strong emphasis on the classification of persistence diagrams. It has been pre-published in [HBL23] and is a joint work with Gilles Blanchard and Clément Levrard. Finally, in Section V, we break free of the computational burden and the mono-persistence constraint of persistence diagrams by computing Euler characteristic type tools and their integral transforms to perform multiple statistical learning tasks. This section is joint work with Vadim Lebovici, pre-published in [HL23].

II Introduction (Français)

L'explosion du nombre de données disponibles est au coeur d'une véritable révolution scientifique et sociétale. De la génération de texte à la classification d'images en passant par la prédiction de tendance sur des séries temporelles, les sciences des données suscitent un intérêt grandissant dans le développement d'outils mathématiques appropriés. Un des principaux défis actuels en apprentissage automatique est le traitement de données vivant dans des espaces de grande dimension. Remettant en cause toute intuition, comme illustré dans le premier chapitre de [Gir14], ces jeux de données font généralement échouer les méthodes d'apprentissage classiques. Cependant, certaines données peuvent être vues comme ayant une structure beaucoup plus simple. En particulier, *l'hypothèse de variété* affirme que beaucoup de jeux de données réels vivent en réalité sur des structures non-Euclidiennes de faible dimension. Un exemple classique est celui des images naturelles : bien que la dimension effective d'une donnée soit son nombre de pixels, les nombreuses contraintes (grandes zones constantes, bords, coins...) réduisent fortement le nombre de degrés de liberté. Ainsi, les images d'un jeu de données spécifique sont souvent considérés comme vivant sur une variété de basse dimension.

Le développement d'outils permettant l'analyse de tels jeux de données "géométriques" est le credo de l'analyse topologique de données (TDA pour *Topological Data Analysis*). La TDA permet l'extraction d'information topologique et géométrique de divers jeux de données en utilisant des outils issus de la topologie algébrique. La TDA et ses applications permet de répondre aux deux objectifs suivants :

- Construire des descripteurs topologiques sur les données afin d'effectuer une tâche d'apprentissage automatique telle que de la classification ou de la régression, voir [CM21].
- Aider à la compréhension qualitative des données via une approche topologique, voir par exemple [Hes20] et [RB19].

Trouvant ses origines dans les travaux sur l'homologie persistante de [ELZ00] et [CZCG04], l'analyse topologique des données a rencontré un grand succès grâce à son large champ d'applications, en particulier en médecine [RYB+20, FM22, ACC+21], en neurologie [KDS+18], biologie [IOH20, RB19], science des matériaux [LBD+17, HNH+16], cosmologie [PEVdW+17], et plus récemment en théorie musicale [AAPL22, MBP22].

Un des objets les plus omniprésents en TDA est le *diagramme de persistance*, qui résume toute l'information topologique présente dans les données. Ce descripteur est le fil rouge de cette dissertation et nous allons étudier son utilisation en apprentissage automatique, et comment dépasser ses limites pratiques et théoriques. Cette thèse est basée sur les trois articles de recherche suivants :

- L'article [HBB+22], en collaboration avec Gilles Blanchard, Krishnakumar Balasubramanian, Clément Levrard et Wolfgang Polonik, où l'on étudie l'utilisation des diagrammes de persistance pour promouvoir la régularité de la fonction estimée dans un problème de régression.
- La pré-publication [HBL23], en collaboration avec Gilles Blanchard et Clément Levrard, où l'on étudie les propriétés structurelles des diagrammes de persistance, et propose une approche et des outils novateurs en classification de mesures afin d'extraire de l'information sur les diagrammes et classifier des données.
- La pré-publication [HL23], en collaboration avec Vadim Lebovici, où l'on étudie des descripteurs qui permettent d'obtenir une meilleure performance que les diagrammes de

persistance tout en ayant un coût de calcul plus faible et en préservant ses propriétés d'interprétabilité.

Après une brève introcution à l'analyse topologique des données et ses concepts clés en Section II.1, nous exposerons les idées et contributions principales de chacun de ces articles dans les Sections II.2, II.3 et II.4. Chaque section principale de ce manuscrit correspond ensuite à un de ces trois articles.

II.1 Une brève introduction à l'analyse topologique des données et aux diagrammes de persistance

On se référera aux livres [EH22] et [BCY18] pour une introduction détaillée à la topologie computationnelle et à l'analyse topologique des données. Nous présentons ici les concepts et notions principales en TDA.

II.1.1 Homologie simpliciale et diagrammes de persistance

Avant de présenter les fondations de la théorie de la persistance, commençons par quelques notions sur les complexes simpliciaux afin de se construire une intuition sur la nature topologique des données dans un cas discret :

Définition II.1. Un *complexe simplicial abstrait (fini)* $\mathcal{K}$ est une collection finie d'ensembles finis tels que leurs sous-ensembles appartiennent également à $\mathcal{K}$. Un élément $\sigma \in \mathcal{K}$ est appelé un *simplexe*, et les sous-ensembles de σ sont appelés les *faces* de σ. La dimension d'un complexe simplicial est la plus grande dimension de l'un de ses simplexes.

On peut montrer, voir Chapitre III de [EH22], qu'en associant chaque sommet abstrait à un point de $\mathbb{R}^{2d+1}$, tout complexe simplicial de dimension d peut être réalisé géométriquement dans $\mathbb{R}^{2d+1}$, où un *simplexe de dimension* k est l'enveloppe convexe de $k+1$ points affinement indépendants. Un simplexe de dimension 0 (resp. 1, 2, 3) est appelé un sommmet ou un noeud (resp. une arête, un triangle, un tétrahèdre). Un des principaux exemples de complexe simplicial construit sur un nuage de points est le complexe de Čech :

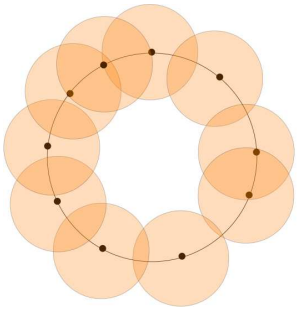
Définition II.2. Soit $\mathbb{X} \subseteq \mathbb{R}^d$ un ensemble de points fini. Le *complexe de Čech à l'échelle* $t \geq 0$ est le complexe simplicial $\check{\mathcal{C}}(\mathbb{X}, t)$ tel que pour tout $(x_0, \dots, x_k) \in \mathbb{X}^{k+1}$, le simplexe $\{x_0, \dots, x_k\}$ appartient à $\check{\mathcal{C}}(\mathbb{X}, t)$ si l'intersection des boules fermées $\cap_{l=0}^{k} \overline{B}(x_l, t)$ est non-vide.

Sur la Figure 10, empruntée à [Wik23], on illustre la construction du complexe de Čech pour un échantillon sur un cercle. On peut voir sur la Figure 10b que ce complexe simplicial ressemble à un cercle simplifié, puisqu'il possède une composante connexe et un seul grand cycle.

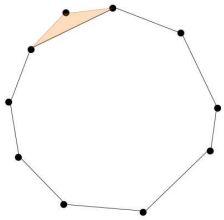
On peut voir sur la Figure 10b que ce complexe simplicial ressemble à un cercle simplifié, puisqu'il possède une composante connexe et un seul grand cycle.

Questions II.3. Se posent alors naturellement deux questions :

- Comment caractériser algébriquement un cycle et le fait qu'ici, le complexe simplicial est connexe et possède un seul cycle ?
- Ici, l'union des boules centrées en chaque point à la même topologie qu'un cycle pour une gamme de paramètres d'échelle t, y compris celui choisi en Figure 10a. Néanmoins, si l'échelle choisie n'est pas adaptée, on peut avoir une union de boules qui n'a pas la



(a) Boules centrées en chaque point



(b) Complexe de Čech correspondant

Figure 10: Construction du complexe de Čech pour un échantillon de points sur un cercle.

topologie souhaitée, par exemple en observant plusieurs composantes connexes. Comment choisir le rayon des boules, c'est à dire le paramètre d'échelle du complexe de Čech, sans aucun a priori sur les données ?

Afin de répondre à la première question, il nous faut introduire quelques notions d'*homologie simpliciale*. Commençons par les définitions d'une k-chaîne et de l'opérateur de bord.

Définition II.4. Soit $\mathcal{K}$ un complexe simplicial. Une *k-chaine simpliciale* est une somme finie (formelle) $\sum_{i=1}^{N} c_i\sigma_i$ où chaque σ_i est un k-simplexe de $\mathcal{K}$ et $c_i \in \mathbb{Z}/2\mathbb{Z}$. L'ensemble des $k-$chaines possède une structure de groupe et est noté C_k.

Définition II.5. L'*opérateur de bord* $\partial_k : C_k \to C_{k-1}$ est le morphisme de groupes tel que:

$$\partial_k(\sigma) = \sum_{i=0}^{k} \{v_0, \ldots, \hat{v_i}, \ldots, v_k\},$$

où $\sigma = \{v_0, \ldots, v_k\}$, et le simplexe de dimension $(k-1)$ $\{v_0, \ldots, \hat{v_i}, \ldots, v_k\}$ est la face de σ obtenue en ôtant le sommet v_i. On définit également les deux sous-groupes de C_k suivants : le groupe des cycles $Z_k = \text{Ker } \partial_k$ et le groupe des frontières $B_k = \text{Im } \partial_{k+1}$.

Par exemple, le bord d'un triangle est la somme de ses trois côtés. On peut aisément vérifier que $\partial_{k+1}\partial_k = 0$, i.e. que le bord du bord d'une chaîne est toujours égal à 0. Ceci implique que B_k est un sous-groupe de Z_k. Intuitivement, un " trou de dimension k" est un cycle de dimension k, qui n'est pas le bord d'un complexe simplicial. On peut alors rigoureusement définir la notion de trou topologique :

Définition II.6. Le *$k$ème groupe d'homologie* H_k de $\mathcal{K}$ est le groupe abélien quotient $H_k(\mathcal{K}) = Z_k/B_k$. Son rang $\beta_k = \text{rank } (H_k(\mathcal{K}))$ est appelé *$k$ème nombre de Betti*.

Intuitivement, le $k$ème nombre de Betti β_k compte le nombre de k-trous pour $k \geq 1$, et β_0 est le nombre de composantes connexes. Revenons à l'exemple de Figure 10b. Il y a une longue chaîne qui n'est pas une frontière. Ainsi, $\beta_1 = 1$. Puisque le complexe simplicial est connexe, on a $\beta_0 = 1$. L'ensemble des nombres de Betti d'un complexe simplicial apparaît donc comme une signature de la topologie d'un complexe simplicial. Dans cette logique, on peut définir un autre invariant appelé *caractéristique d'Euler* :

Définition II.7. La *caractéristique d'Euler* d'un complexe simplicial $\mathcal{K}$ de dimension d est l'entier relatif $\chi(\mathcal{K})$ défini par

$$\chi(\mathcal{K}) = \sum_{\sigma \in \mathcal{K}} (-1)^{\dim \sigma} = \sum_{k=0}^{d} (-1)^k \beta_k(\mathcal{K}).$$

La caractéristique d'Euler est moins informative que la collection de tous les β_k pour $k \in \{0, \dots, d\}$. Cependant, son expression obtenue en comptant les simplexes en fait un invariant beaucoup plus simple à calculer que les nombres de Betti. Ainsi, nous ferons un usage intensif de la caractéristique d'Euler en Section V et démontrerons qu'il s'agit d'un descripteur extrêmement puissant en analyse de données.

On peut maintenant traiter la seconde question parmi les Questions II.3, ce qui est le but de l'*homologie persistante*. Intuitivement, la meilleure manière de choisir la taille des boules pour la construction du complexe de Čech est de ne pas la choisir, ou plutôt de considérer toutes les tailles possible. Plus formellement, on peut commencer par définir une *filtration* d'un complexe simplicial :

Définition II.8. Soit $\mathcal{K}$ un complexe simplicial fini et soit $f : \mathcal{K} \to \mathbb{R}$ une fonction croissante au sens de l'inclusion : $f(\sigma) \leq f(\tau)$ lorsque σ est une face de τ. Pour tout $a \in \mathbb{R}$, le sous-ensemble de niveau $\mathcal{K}(a) = f^{-1}((-\infty, a])$ est un sous-complexe simplicial de $\mathcal{K}$. En considérant toutes les valeurs possibles de f, on a une famille emboîtée de complexes simpliciaux :

$$\varnothing = \mathcal{K}_0 \subset \mathcal{K}_1 \subset \ldots \subset \mathcal{K}_n = \mathcal{K},$$

appelée *filtration*, où $a_0 = -\infty < a_1 < a_2 < \ldots < a_n$ sont les valeurs prises par f sur les simplexes de $\mathcal{K}$.

Le complexe de Čech défini précédemment sur un nuage de points $\mathbb{X}$ défini naturellement une filtration sur le complexe simplicial complet $\mathcal{K} = 2^{\mathbb{X}}$:

$$\varnothing \subset \check{\mathcal{C}}(\mathbb{X}, 0) = \mathbb{X} \subset \check{\mathcal{C}}(\mathbb{X}, t_1) \subset \check{\mathcal{C}}(\mathbb{X}, t_2) \subset \ldots \subset \check{\mathcal{C}}(\mathbb{X}, t_n) = \mathcal{K},$$

où $0 < t_1 < t_2 < \ldots < t_n$ sont les rayons de boules auxquels on observe un changement de topologie du complexe de Čech. Cette filtration est simplement appelée *filtration de Čech* et notée $\check{\mathcal{C}}(\mathbb{X})$. Le complexe simplicial $\mathcal{K}$ étant fini, le nombre de rayons critiques t_i pour la topologie du complexe de Čech est fini. On peut de plus montrer, voir [BE17], que les t_i sont des rayons de boules circonscrites aux simplexes de $\mathcal{K}$.

Ici, l'idée est donc de considérer une famille imbriquée de complexes simpliciaux et de prendre en compte la création et la destruction de composantes topologiques, c'est à dire l'évolution des nombres de Betti. On peut formaliser cela ainsi : pour $i \leq j$, l'application inclusion canonique $\mathcal{K}_i \to \mathcal{K}_j$ induit un morphisme de groupes $f_k^{i,j} : H_k(\mathcal{K}_i) \to H_k(\mathcal{K}_j)$ sur les groupes d'homologie pour tout k, donnant naissance à une suite de groupes d'homologies imbriqués :

$$0 = H_k(\mathcal{K}_0) \to H_k(\mathcal{K}_1) \to \ldots \to H_k(\mathcal{K}_n) = H_k(\mathcal{K}).$$

On peut alors définir l'*homologie persistante* :

Définition II.9. Les $k$ème *groupes d'homologie persistante* sont les images des morphismes définis ci-dessus : $H_k^{i,j} = \text{im } f_k^{i,j}$. Comme pour l'homologie, on définit le $k-$ème *nombre de Betti persistant* par $\beta_k^{i,j} = \text{rank } H_k^{i,j}$.

Comme son nom l'indique, le $k$ème nombre de Betti persistant $\beta_k^{i,j}$ correspond au nombre de k-trous qui *persistent* entre $\mathcal{K}_i$ et $\mathcal{K}_j$. Etant donné une classe d'éléments $\gamma \in H_k(\mathcal{K}_i)$, on dit que cette classe est *née* en $\mathcal{K}_i$ si $\gamma \notin H_k^{i-1,i}$. De même, l'instant de *mort* de γ (éventuellement infini) est le plus petit indice j tel que $f_k^{i,j}(\gamma) \in H_k^{i-1,j}$. Nous avons désormais tous les outils à notre disposition pour définir l'object central de cette dissertation.

Définition II.10. Le $k$ème *diagramme de persistance* est le multi-ensemble de $\overline{\mathbb{R}}^2$ des coordonnées (naissance, mort) pour chaque classe γ qui a existé à un moment donné de la suite d'homologie persistante.

Ce multi-ensemble peut aisément être vu comme une mesure discrète ce qui sera au coeur de la Section IV. De plus, remarquons que le nombre de Betti persistant $\beta_k^{i,j}$ correspond au nombre de points dans le quadrant infini avec un angle en (a_i, a_j) en haut à gauche du diagramme de persistance. Sur la Figure 11, adaptée de [RCL+21], on peut voir la construction d'un diagramme de persistance d'un complexe de Čech d'un nuage de points.

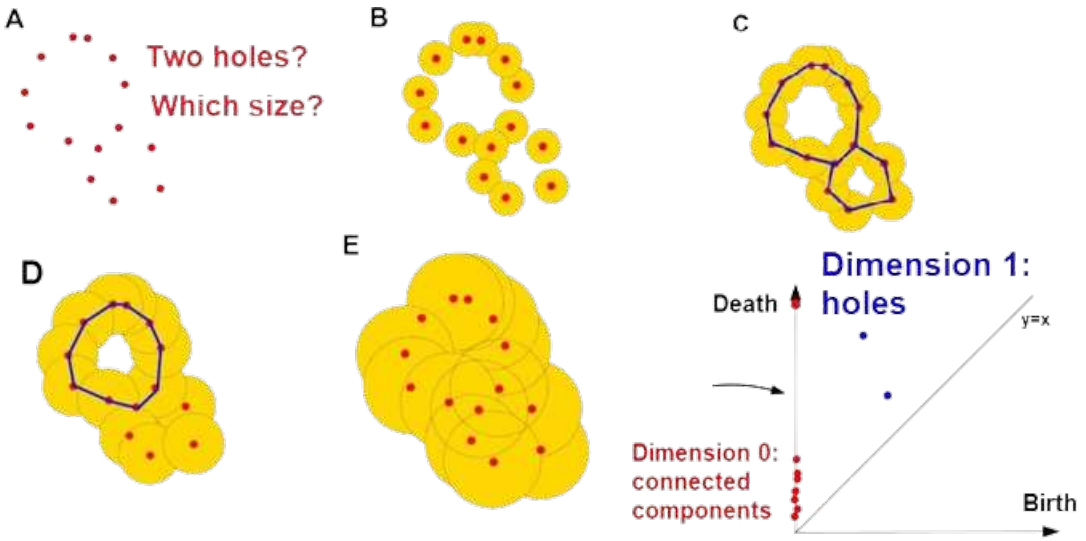


Figure 11: Diagrammes de persistance du complexe de Čech sur un nuage de points.

Ici, les points sont échantillonnés sur un bouquet de deux cercles. La topologie d'un nuage de points brut est simplement constituée d'autant de composantes connexes qu'il y a de points. Afin de construire le complexe de Čech de ce nuage de points, on centre des boules en chaque point et on fait varier leur rayon de 0 à $+\infty$. L'une après l'autre, les boules de points consécutifs vont se toucher, créant des arêtes dans la filtration de Čech et fusionnant ainsi des composantes connexes. Topologiquement, une fusion correspond à la mort d'une des deux composantes. Ainsi, le diagramme de persistance pour la 0-homologie contient uniquement des points de coordonnées $(0, d_i)$ ainsi qu'un point de coordonnées $(0, \infty)$ puisque lorsque les boules sont suffisamment grandes (Figure 11, E), l'union des boules reste connexe. Quant aux cycles, l'union des boules ne contient aucun cycle pour des petits rayons (Figure 11, B). Lorsque le rayon devient suffisamment grand (Figure 11, C), on peut voir que deux cycles apparaissent dans le complexe de Čech. Le rayon critique conduisant à l'apparition d'un cycle est son *temps de naissance*. Finalement, quand le rayon devient suffisamment grand

(Figure 11, D et E), la triple intersection de boules crée des triangles dans le complexe de Čech qui remplissent les cycles. Le rayon critique correspondant est le *temps de mort* du cycle. On a ainsi deux points dans le diagramme de persistance correspondant à la 1-homologie. En plus de caractériser la topologie d'un nuage de points (ici le fait qu'il possède deux cycles), les diagrammes de persistance donnent également de l'information sur la taille géométrique des trous. De plus, on peut extraire de l'information sur l'échantillonnage et les distances entre points du diagramme de persistance de dimension 0.

Le succès de l'analyse topologique de données vient entre autres de la possibilité de comparer des diagrammes de persistance. Une des approches les plus populaires à cette fin se fait via la *distance bottleneck* définie ainsi :

Définition II.11. Soit $\Delta = \{(x, x) | x \in \mathbb{R}\}$ la diagonale de $\mathbb{R}^2$.

La *distance bottleneck* d_B entre deux diagrammes de persistance D et D' est :

$$d_B(D, D') = \inf_{\eta : D \cup \Delta \to D' \cup \Delta} \sup_{x \in D \cup \Delta} \|x - \eta(x)\|_\infty,$$

où l'infimum est pris sur toutes les bijections η entre $D \cup \Delta$ et $D' \cup \Delta$.

On illustre l'appariement optimal entre deux diagrammes en Figure 12, inspirée de [Cha23]. La distance bottleneck est ici la longueur de la plus longue flèche (en norme infinie). On se référera à [DL21] pour un traitement détaillé de la structure de l'ensemble des diagrammes de persistance munis de métriques similaires. Une des forces des diagrammes de persistance est leur propriété de stabilité, voir [CSEH07], qui affirme qu'une petite perturbation dans les données induira une petite perturbation dans les diagrammes, mesurée en termes de distance bottleneck. Nous citerons le résultat correspondant en section II.1.2

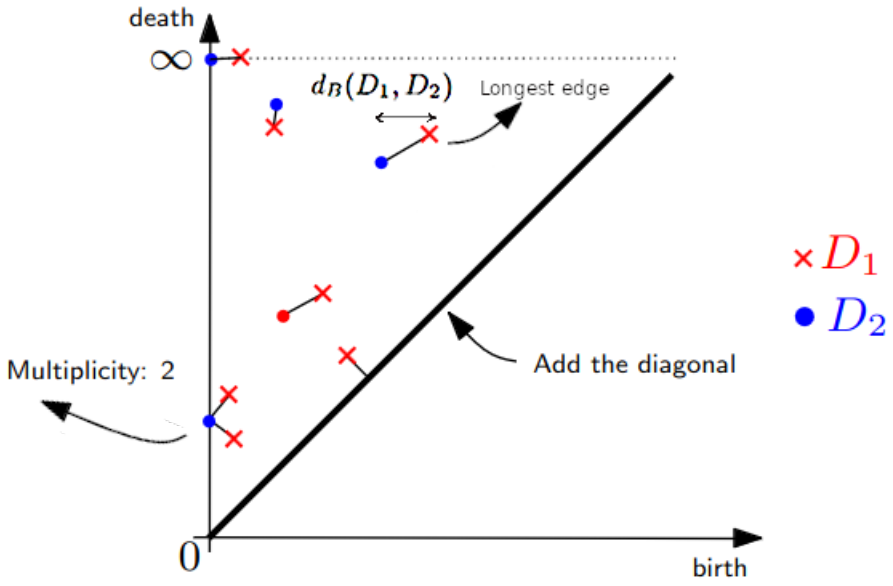


Figure 12: Appariement optimal entre deux diagrammes de persistance ayant des nombres de points différents

Grâce à son approche multi-échelle, les diagrammes de persistance contiennent énormément d'information de nature topologique sur les données. Ces descripteurs peuvent alors être

utilisés pour diverses tâches d'apprentissage telles que la classification, la régression, ou le partitionnement. On peut mentionner l'étude suivante [HMR21] qui traite de l'applications des diagrammes de persistance à l'apprentissage automatique en pratique. En raison de sa représentation en tant qu'ensemble de $\overline{\mathbb{R}}^2$ (ou de mesures discrètes), les diagrammes de persistance ne peuvent être utilisés tels quels comme entrée d'algorithmes classiques d'apprentissage machine. Une stratégie usuelle est de transformer les diagrammes en éléments d'espaces de Banach. Citons entre autres les courbes de Betti [RSL20], les images persistantes [AEK+17], les landscapes [B+15], et plus récemment des techniques de vectorisation de mesure, [RCL+21] et des méthodes faisant appel à des réseaux de neurones [CCI+20, RCB21]. Nous verrons en Section IV une méthode permettant de faire de l'apprentissage statistique directement dans l'espace des mesures sans aucune étape de vectorisation.

D'un point de vue computationnel, les complexes simpliciaux, l'homologie persistante et les diagrammes de persistance peuvent être calculés en utilisant la librairie `Gudhi`, [MBGY14]. L'algorithme originel permettant le calcul de la persistance peut être trouvé dans [ZC04]. Le calcul d'un diagramme de persistance pour un complexe simplicial $\mathcal{K}$ a une complexité au pire cas en $\mathcal{O}(|\mathcal{K}|^\omega)$ où $2 \leq \omega < 2.373$ est l'exposant pour la multiplication de matrices, voir [MMS11]. Bien que le calcul des diagrammes de persistance paraisse raisonnable, les algorithmes actuels se comportent mal en grande dimension. En effet, pour un complexe géométrique typique, le nombre de simplexes croît exponentiellement avec la dimension.

Enfin, les diagrammes de persistance peuvent bien sûr être calculés dans un cadre beaucoup plus général que celui de complexes de Čech sur des nuages de points. Dans la Section V, on explorera d'autres filtrations sur les nuages de points, tandis que dans les Sections IV et V, on verra comment appliquer cela à de l'apprentissage sur des graphes. Cette méthodologie peut également être appliquée à l'étude d'images et de volumes 3D en considérant des *complexes cubiques* à la place de complexes simpliciaux. Cette approche a en particulier permis le développement de méthodes novatrices en analyse d'images médicales, par exemple dans [ACC+21] ou encore [JKN20]. De plus, la théorie de l'homologie persistante va bien plus loin que l'étude de complexes et de filtrations finis. Elle peut en particulier être appliquée à des suites d'espaces topologiques imbriqués, toujours dans le but d'extraire des temps de naissance et de morts de composantes topologiques. Un des exemples principaux est l'étude des sous-niveaux de fonctions de Morse. Cet exemple motive la Section III de ce travail, et nous allons maintenant exposer certains de ces aspects en Section II.1.2.

II.1.2 Persistance pour les fonctions de Morse

Commençons cette section par la définition des fonctions de Morse et en énonçant un résultat sur la topologie de leurs sous-niveaux.

Définition II.12. Soit $f : \Omega \to \mathbb{R}$ une fonction $\mathcal{C}^2$ où Ω est inclus dans $\mathbb{R}^d$.

- Un *point critique* de f est un point $x \in \Omega$ tel que $\nabla f(x) = 0$. Son image $f(x)$ est appelée *valeur critique*.
- Soit x_0 un point critique de f. Son *indice* est le nombre de valeurs propres négatives de la Hessienne de f en x_0.

Définition II.13. Une fonction f définie sur une variété différentielle est une *fonction de Morse* si elle est $\mathcal{C}^2$ et que tous ses points critiques sont non-dégénérés, i.e. telle que la matrice Hessienne en chaque point critique est non-dégénérée.

L'ensemble des fonctions de Morse est un ouvert dense de l'ensemble des fonctions continues. Une des propriétés principales des fonctions de Morse est résumée dans le théorème suivant, adaptée des Théorèmes 3.1 et 3.2 de [Mil63]:

Théorème II.14. *Soit f une fonction de Morse sur une variété lisse $\mathcal{M}$ et soit $\mathcal{M}^a$ le sous-ensemble de niveau $f^{-1}((-\infty, a])$.*

- *Soit $a < b$. Supposons qu'il n'y ait pas de valeur critique entre a et b. Alors $\mathcal{M}^a$ et $\mathcal{M}^b$ sont difféomorphes et $\mathcal{M}^b$ se rétracte par déformation sur $\mathcal{M}^a$.*
- *Soit p un point critique de f d'indice s tel que $f(p) = q$. Supposons de plus qu'il n'y ait pas d'autre point critique p' tel que $f(p') = q$. Alors, pour ε suffisamment petit, $\mathcal{M}^{q+\varepsilon}$ a le même type d'homotopie que $\mathcal{M}^{q-\varepsilon}$ auquel on a attaché une s-anse.*

Ce théorème illsutre que les changements de topologie des sous-ensembles de niveau d'une fonction de Morse se produisent aux valeurs critiques de la fonction. Par analogie avec l'homologie simpliciale, considérons l'évolution des groupes d'homologie des sous-niveaux $f^{-1}((-\infty, t])$ lorsque t parcourt $\mathbb{R}$ de $-\infty$ à $+\infty$, et notons dans un diagramme de persistance les valeurs auxquelles les composantes topologiques naissent et meurent. On propose d'illustrer cela en Figure 13, inspirée de [CM21], où l'on considère la fonction hauteur sur un "tore perturbé" avec ses 0 et 1-diagrammes de persistance, respectivement en rouge et bleu.

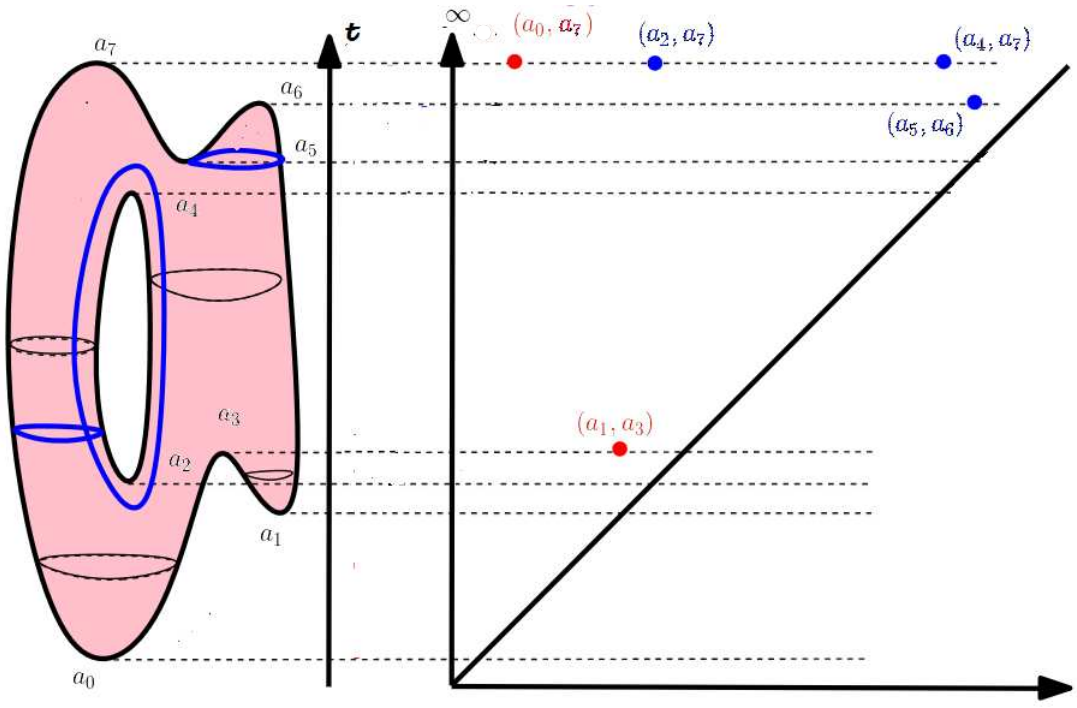


Figure 13: Filtration par les sous-niveaux d'un tore perturbé et diagrammes de persistance correspondant

Considérons un plan d'équation $z = t$ et faisons évoluer t de $-\infty$ à $+\infty$. Les sous-niveaux sont tous vides avant d'atteindre le minimum global a_0. Ensuite en a_0, une composante connexe naît. Une autre composante connexe est créée au minimum local a_1, et les deux composantes connexes fusionnent finalement au point-selle a_3. Par la convention de la *règle*

de l'ancien, on considère que la composante la plus jeune meurt en premier, donnant lieu à un point de coordonnées (a_1, a_3) dans le diagramme de persistance. A partir de a_2, les sous-niveaux contiennent un 1-cycle, et en a_4, un deuxième cycle est créé. La figure a désormais le type d'homologie d'un tore. Un troisième cycle est créé au point-selle a_5, qui persiste seulement jusqu'au maximum local a_6, donnant un point de coordonnées (a_5, a_6) dans le 1-diagramme de persistance. Après a_7, les sous-niveaux ont de nouveau la même topologie que le tore. Ils restent connexes pour toujours et la composante connexe créée en a_0 ne meurt jamais. De même, les cycles créés en a_2 et a_7 persistent pour toujours. Par convention, afin d'éviter d'avoir des coordonnées infinies, on impose le temps de mort au maximum global de la fonction, à savoir a_7.

Ici, on peut voir l'impact du Théorème II.14 pour le calcul des diagrammes de persistance puisque les coordonnées de tous les points sont des valeurs critiques de la fonction hauteur. Plus précisément, pour cette fonction de $\mathbb{R}^2$ dans $\mathbb{R}$, les minima créent systématiquement les composantes connexes tandis que les maximas tuent systématiquement les cycles. Quant aux points-selles, ils peuvent soit fusionner deux composantes connexes, soit donner naissance à un cycle.

Nous disposons désormais d'un multi-ensemble qui décrit une fonction de Morse et ses points critiques d'un point de vue topoloique. Etant donné les applications en apprentissage statistique visées par ce manuscrit, une question naturelle est le comportement de ce descripteur lorsque l'on ajoute du bruit à la fonction. Si le bruit est borné par ε, on peut s'attendre à ce que chaque point critique soit déplacé d'au plus ε. Ainsi, de nombreux points critiques qui ne persistent pas au-delà de ε vont apparaître. Nous précisons cette observation dans le résultat de stabilité suivant, issu de [CSEH07], en terme de la distance bottleneck introduite en Définition II.11. On note D_f le diagramme de persistance des sous-niveaux d'une fonction de Morse f.

Théorème II.15. *Soit $\mathcal{M}$ un espace topologique et soit f et g deux fonctions de Morse de $\mathcal{M}$ vers $\mathbb{R}$. On a alors :*

$$d_B(D_f, D_g) \leq \|f - g\|_\infty.$$

Disposer d'une métrique possédant de telles propriétés de stabilité paraît très intéressant dans un contexte d'analyse de données car elle démontre une certaine robustesse à un bruitage borné. Dans notre contexte, on aimerait étudier le *débruitage* d'une fonction en utilisant de tels régulariseurs topologiques. A cette fin, on introduit la *persistance* d'une fonction:

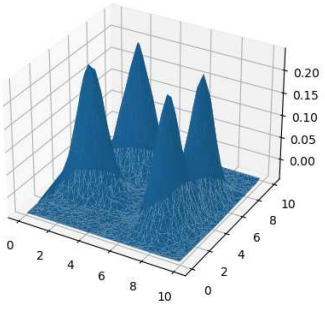
Définition II.16.
- Pour un point (b, d) d'un diagramme de persistance, sa *persistance* est égale à son temps de vie $d - b$.
- La *persistance* d'un diagramme D est la somme de tous les temps de vie de chacun de ses points :

$$\mathrm{Pers}(D) = \sum_{(b,d)\in D} (d - b).$$

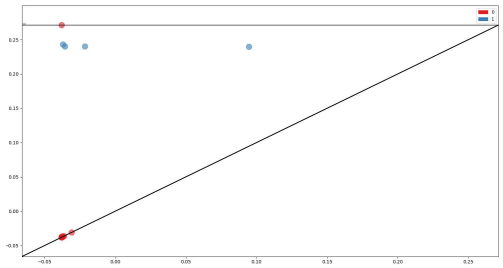
S'il n'y a pas d'ambiguïté, on notera $\mathrm{Pers}_k(f)$ la persistance du $k$ème diagramme de persistance des sous-niveaux d'une fonction f et $\mathrm{Pers}(f) = \sum_k \mathrm{Pers}_k(f)$ sa *persistance totale*.

On illustre en Figure 14 le comportement de la persistance et de la distance bottleneck lorsque l'on ajoute un bruit gaussien à des mesures d'une fonction lisse. Plus précisément, considérons des observations $Y_i = f^\star(X_i) + \varepsilon_i$ où les points $(X_i)_{i=1}^{1000}$ sont échantillonnés

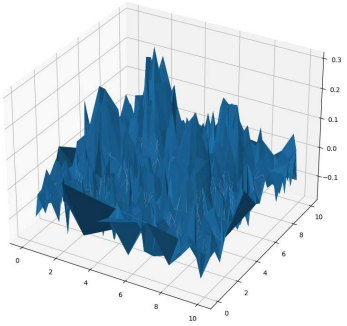
uniformément sur le carré unité, $f^\star$ est une somme de quatre gaussiennes représentée en Figure 14a, et $\varepsilon \sim \mathcal{N}(0, \sigma I_n)$. On affiche une interpolation des fonctions lisses et bruitées au carré unité en Figures 14a and 14c. Lorsque l'on ajoute du bruit à chaque mesure, de nombreux points critiques avec une faible persistance sont ajoutés et envoyés sur la diagonale lors du calcul de la distance bottleneck, illustrant ainsi le Théorème II.15 avec grande probabilité. La distance bottleneck mesure ici le bruit ajouté à une fonction. Cependant, elle nécessite la connaissance de la vraie fonction de régression $f^\star$. De même, la 0 et 1-persistance mesurent la quantité de bruit ajouté à une fonction. Il existe également des résultats de stabilité pour la persistance qui font intervenir le nombre total de points dans le diagramme, sur lequel on n'a qu'un contrôle restreint. On énoncera ce résultat précis dans le Lemme III.2.



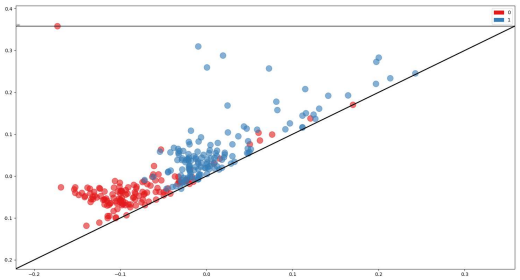
(a) Vraie fonction



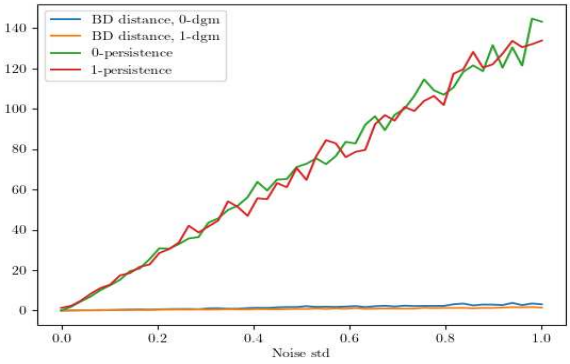
(b) Diagramme de persistance correspondant



(c) Fonction bruitée



(d) Diagramme de persistance correspondant



(e) Distances bottleneck et persistance en fonction de l'écart-type du bruit

Figure 14: Stabilité de la distance bottleneck et de la persistance à un bruit Gaussien

On dispose désormais de tous les outils pour passer au coeur de la première contribution de cette dissertation, où l'on s'intéresse à l'utilisation de la persistance totale pour résoudre un problème de régression.

II.2 Imposer de la régularité grâce aux diagrammes de persistance

La persistance totale capture les petites oscillations d'une fonction et est ainsi particulièrement indiquée comme un régularisateur dans un contexte de régression. Cette observation est au coeur de la Section III, publiée dans [HBB+22]. L'objectif ici est d'utiliser la persistance totale comme un terme de pénalité dans le but d'annuler le bruit des observations, résultant en une prédiction plus lisse et plus précise. Ce travail se focalise sur un problème de régression où l'on observe des données sur une variété.

II.2.1 Régularisation topologique

On considère un problème de régression sur une variété compacte $\mathcal{M}$. On suppose que les données sont des points $(X_i)_{i=1}^n$ échantillonnés uniformément et indépendamment sur $\mathcal{M}$ et que l'on observe des étiquettes réelles :

$$Y_i = f^\star(X_i) + \varepsilon_i,$$

où $(\varepsilon_i)_{1\leq i\leq n}$ sont des variables aléatoires indépendantes entre elles et indépendantes de tous les X_i, identiquement distribuées, sous-gaussiennes et de moyenne nulle. Notre but est de retrouver la vraie fonction $f^\star$ dans une optique de débruitage ou de prédiction étant donné une nouvelle observation sur la variété.

Commençons par considérer une base de fonctions $(\Phi_i)_{i\geq 0}$ adaptée à la variété, appelées les fonctions propres de l'*opérateur de Laplace-Beltrami*, voir [Ros97]. Cette base peut être vue comme une généralisation de la base de Fourier à un espace fonctionnel sur une variété compacte. En pratique, on a rarement accès aux fonctions propres $(\Phi_i)_{i\geq 0}$. Cependant, on peut les approcher en utilisant le spectre de la matrice Laplacienne d'un graphe construit sur les données, voir [Chu97]. Afin de trouver les meilleurs coefficients pour décomposer $f^\star$ sur la base des p premières fonctions propres de l'opérateur de Laplace-Beltrami (ou du graphe Laplacien), on minimise le critère suivant :

$$\mathcal{L}(\theta) = \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\theta_j\Phi_j(x_i)\right)^2 + \mu\Omega(\theta),$$

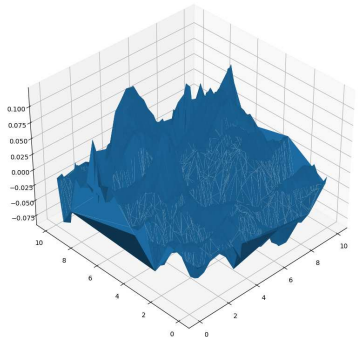
où Ω est un terme de pénalité visant à empêcher les phénomènes de surinterprétation et est calibré par le réel μ.

Parmi les exemples de pénalités classiques, on peut citer la régularisation L^1 ou Lasso (voir [Bv11] pour un traitement exhaustif) la régularisation L^2 ou Ridge, [HK70], ainsi que la pénalisation par variation totale, [HR16], qui est celle qui se rapproche le plus de notre travail, puisqu'elle vise à pénaliser les fortes oscillations. Ici, on considère les deux pénalités suivantes :

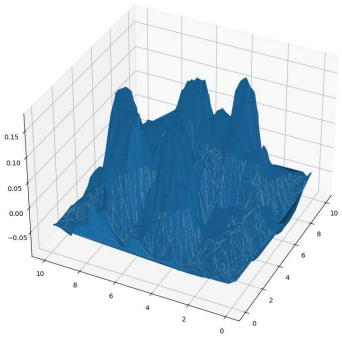
- $\Omega_1(\theta) = \sum_{j=1}^{p}|\theta_j|\mathrm{Pers}(\Phi_j)$ est une pénalité de type Lasso pondéré qui s'implémente très facilement et possède un coût de calcul très modéré. Cette pénalité permet d'effectuer une sélection de variables en privilégiant les fonctions propres avec une persistance restreinte, c'est à dire celles qui n'oscillent pas trop.

- $\Omega_2(\theta) = \text{Pers}\left(\sum_{j=1}^{p} \theta_j \Phi_j\right)$ est une pénalité non-convexe, pour laquelle on utilise des techniques introduites dans [CCG+21] afin de minimiser la fonction de perte $\mathcal{L}$. Comme illustré en Figure 14, cette pénalité va permettre de faire du débruitage, en éliminant les points peu persistants du diagramme, et offrir ainsi une reconstruction lisse de la fonction observée.

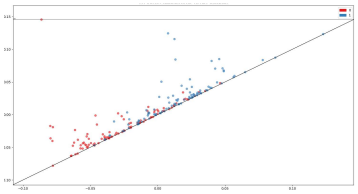
En pratique, on combine ces deux pénalités en sélectionnant un sous-ensemble de fonctions propres grâce à Ω_1, puis en faisant du débruitage en utilisant la pénalité Ω_2. On peut voir les effets de la régularisation par Ω_2 dans la Figure 15 où l'on débruite la fonction de la Figure 14c. La reconstruction à partir d'un échantillon bruité est comparée à une régression sur les fonctions propres du Laplacien avec une pénalité Lasso. On peut voir que la régularisation via Ω_2 est plus lisse, et bien qu'un peu d'information soit perdue au niveau de l'intensité des pics, on parvient néanmoins à retrouver les quatre pics du signal initial. De plus, on peut voir sur les diagrammes de persistance correspondant qu'il y a quatre points loin de la diagonale, correspondant à quatre maxima qui persistent relativement longtemps. En revanche, la pénalisation par Lasso ne parvient à récupérer que deux ou trois pics du signal original et son diagramme de persistance contient plus de points près de la diagonale.



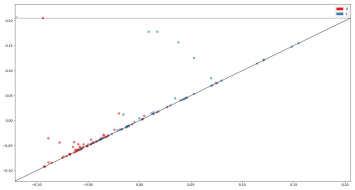
(a) Fonction reconstruite par Lasso



(b) Fonction estimée via la pénalité topologique Ω_2



(c) Diagramme de persistance de l'estimation par Lasso



(d) Diagramme de persistance de l'estimation par la pénalité Ω_2

Figure 15: Reconstruction de la somme de quatre gausiennes.

II.2.2 Liste détaillée des contributions

Les contributions de ce travail sont les suivantes :

(i) On a effectué une analyse approfondie, qualitative et quantitative sur des données simulées et réelles, en comparant les méthodes par pénalisation topologique à des méthodes de régression standard. Notre méthode est comparable en terme de performance à une

régression à noyau avec pénalité Ridge, [Vov13] et surpasse les autres méthodes de régularisation, y compris celles par variation totale.

(ii) On a établi plusieurs résultats oracles pour les pénalités Ω_1 et Ω_2. En particulier, sous l'hypothèse que la fonction de régression se décompose $f^\star = \sum_{j=1}^{p} \theta_j^\star \Phi_j$, le paramètre optimal $\theta^\star$ est approché à une vitesse $O(p/n)$. De plus, on a également un résultat théorique sur la persistance totale de la fonction estimée garantissant que celle-ci est "topologiquement régulière".

(iii) On a établi un résultat négatif sur la capacité de l'espace des fonctions à persistance bornée.

Ce travail contient la première occurence de garanties théoriques pour des modèles de régularisation topologique, qui avaient déjà été utilisés dans la littérature. Bien que les résultats expérimentaux soient prometteurs, le coût computationnel de la méthode reste son principal défaut. En effet, afin de minimiser $\mathcal{L}$ avec la pénalité Ω_2, il faut calculer des centaines de diagrammes de persistance (un par pas de la descente de gradient). De plus, le point de vue adopté ici est extrêmement simpliste en terme d'analyse des diagrammes de persistance. En effet, ceux-ci sont séparés en une composante de signal loin de la diagonale correspondant à peu de choses près au diagramme de la vraie fonction $f^\star$ et en une composante de bruit topologique correspondant au bruit ε sur les observations et qui se manifeste comme un agrégat de points près de la diagonale. On va voir qu'il y a en réalité beaucoup d'information contenue près de la diagonale des diagrammes de persistance, et que celle-ci peut être utilisée à des fins de classification.

II.3 Extraire l'information près de la diagonale

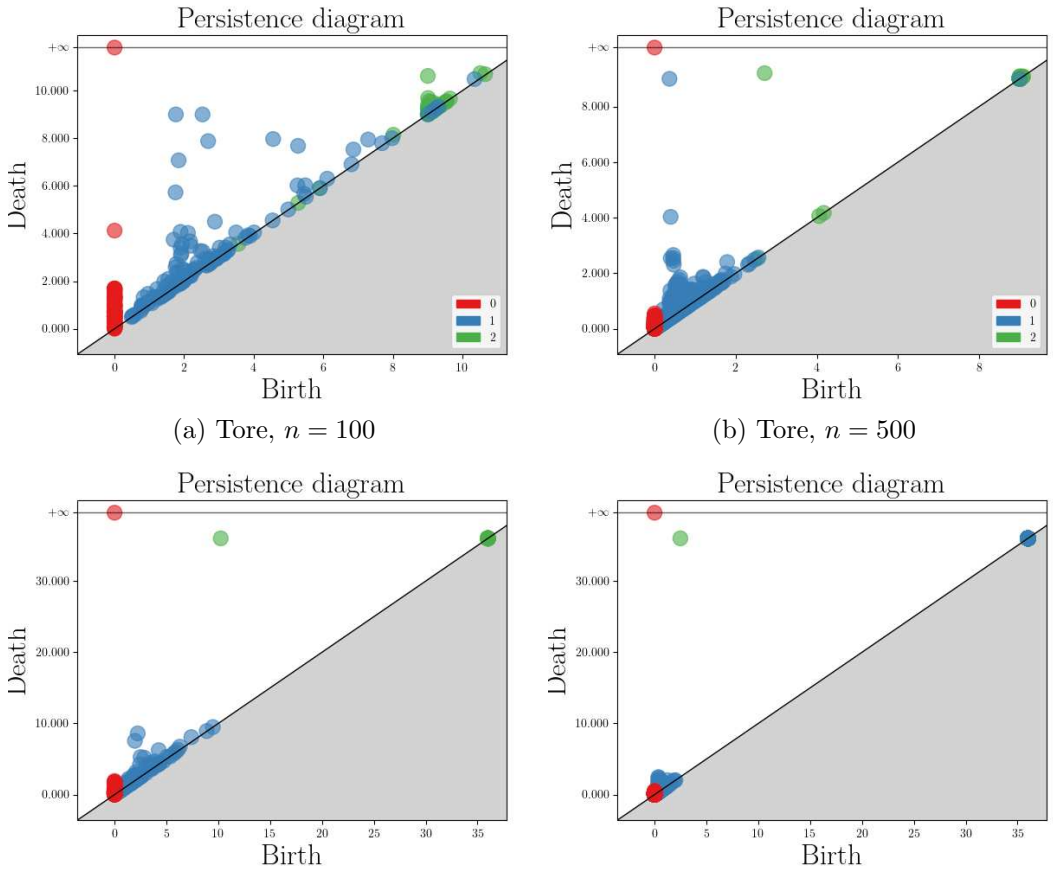
On va maintenant se concentrer sur le cas de diagrammes de persistance de la filtration de Čech sur un nuage de points de $\mathbb{R}^d$. Ce travail a donné lieu à la pré-publication [HBL23]. On considère un problème de classification binaire supervisée où l'on observe des données $D_N = (\mu_1, Y_1), \ldots, (\mu_N, Y_N)$ où μ_i est un diagramme de persistance et Y_i est un label binaire dans $\{0, 1\}$. En pratique, ce type de situation provient d'un problème de classification binaire sur des données brutes (graphes, images, séries temporelles, nuages de points...) transformées en diagrammes de persistance en raison de l'information topologique qu'elles peuvent contenir.

II.3.1 Un exemple motivateur : distinguer un tore d'une sphère

On va essayer de discriminer des nuages de points echantillonnés sur une sphère $\mathbb{S}^2$ ou sur un tore $\mathbb{T}^2$. Une façon de résoudre ce problème de façon invariante par translation et rotation est de regarder les diagrammes de persistance de dimension 1 des complexes de Čech construits sur les données. Sur la Figure 16, on observe des exemples de diagrammes de persistance de complexes de Čech de nuages de points de taille variable, échantillonnés sur une sphère ou un tore.

Ainsi, dès que le nombre de points échantillonnés est suffisamment grand, on peut s'attendre à ce que le 1-diagramme de persistance de la filtration de Čech de l'échantillon sur un tore a deux cycles très persistants tandis que celui de la sphère n'a pas de cycle loin de la diagonale.

En raison de leur structure de multi-ensemble (ou de mesure discrète sur $\overline{\mathbb{R}}^2$), les diagrammes de persistance ne peuvent pas être entrés tels quels dans des algorithmes de classiication supervisée standard. Comme mentionné précédemment, une stratégie usuelle est d'envoyer les diagrammes de persistance dans un espace de Banach [DP19] puis de les transformer en vecteurs. On se propose ici de répondre aux deux questions suivantes :



(a) Tore, $n = 100$ (b) Tore, $n = 500$

(c) Sphère, $n = 100$ (d) Sphère, $n = 500$

Figure 16: Diagrammes de persistance de dimension 0, 1 et 2 de la filtration de Čech sur un échantillon de n points.

- Est-ce possible de contourner cette étape de vectorisation et développer une méthode de classification directement sur les mesures, tout en possédant de bonnes garanties théoriques ?
- Dans l'exemple de classification entre un tore et une sphère (comme dans la Section II.2), on a traité les diagrammes de persistance comme d'une part des points très persistants correspondant à un signal homologique et d'autre part des points peu persistants correspondant à du bruit. Est-il possible de dépasser cette dichotomie, et quel type d'information se cache dans les points près de la diagonale ?

II.3.2 Le problème de classification de mesures

On a développé dans la Section IV une méthode de classification supervisée de mesures s'affranchissant de toute étape de vectorisation. Pour des mesures définies sur un espace métrique compact $\mathcal{X}$, on considère différentes zones de $\mathcal{X}$ et on discrimine entre les classes selon la masse que chaque mesure met en ces zones. Par exemple, dans l'exemple précédent de la sphère et du tore, regarder si les diagrammes de persistance (vus comme mesures) mettent plus de deux points au-dessus d'un certain seuil de mort adéquatement choisi suffit à bien classifier une donnée comme étant issue d'un tore. On propose deux algorithmes afin

d'apprendre les zones de discrimination ainsi que les seuils d'activation correspondant. Ces différentes zones sont ensuite aggrégées en utilisant une procédure de *boosting*. De plus, on fait rentrer ce type de classifieurs dans une classe plus large et on propose des garanties statistiques correspondantes. Plus précisément on considère une classe de fonctions $\mathcal{F}$ sur $\mathcal{X}$, et une classe de fonctions correspondante $\tilde{\mathcal{F}}$ définie sur l'espace $\mathcal{M}(\mathcal{X})$ de mesures $\mathcal{X}$, dont les fonctions sont définies comme :

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

On fait le lien entre des mesures de capacité sur $\tilde{\mathcal{F}}$ et sur $\mathcal{F}$, qui sont en général bien plus simples à calculer pour cette seconde classe. Plus précisément, pour une classe de fonctions générale $\mathcal{F}$, on définit la complexité de Rademacher empirique sur un échantillon $(Z_i)_{i=1}^N$ comme

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

où $(\sigma_1, \ldots, \sigma_N)$ est un vecteur de variables aléatoires indépendantes de Rademacher, i.e. pour tout i, $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$. Cette quantité mesure la corrélation de la classe de fonctions $\mathcal{F}$ avec un vecteur de bruit de Rademacher. Sous des hypothèses assez générales, le risque de prédiction est borné par la complexité de Rademacher avec grande probabilité, voir [SSBD14]. Il est donc préférable pour une famille de classifieurs d'avoir une faible complexité de Rademacher, ce qui signifie que la classe possède de bonnes propriétés de généralisation. Dans notre cas, on a établi la borne supérieure suivante :

Théorème II.17. *Il existe une constante universlle K telle que*

$$\mathcal{R}_N(\tilde{\mathcal{F}}) \leq \frac{K\overline{M}_2\sqrt{VC(\mathcal{F})}}{\sqrt{N}},$$

où $\overline{M}_2 = \left(\frac{1}{N}\sum_{i=1}^N \mu_i(\mathcal{X})^2\right)^{\frac{1}{2}}$, et VC est une autre mesure de capacité, appelée dimension de Vapnik-Chervonenkis. On remarque que la complexité de Rademacher empirique de $\tilde{\mathcal{F}}$ est bornée par une mesure de capacité de la classe $\mathcal{F}$. Dans la Section IV, on propose également une borne inférieure du même ordre de grandeur sur la complexité de Rademacher empirique, et on fait le lien avec des bornes de prédiction pour les classifieurs de mesures proposés.

II.3.3 Résultats asymptotiques sur les complexes de Čech aléatoires

Dans le cas des diagrammes de persistance, l'algorithme précédemment décrit va, dans de nombreux scénarios pratiques, utiliser des zones près de la diagonale pour discriminer entre les deux classes. Cela montre qu'il y a de l'information pertinante contenue dans ce que nous avions jusqu'à présent injustement appelé le *bruit topologique*. On appuie ce constat expérimental par plusieurs garanties théoriques. En effet, dans [BHPW20], les auteurs affirment que les points près de la diagonale contiennent de l'information sur la courbure de l'espace sur lequel les points ont été échantillonnés. En plus du support, ces points près de la diagonale permettent d'extraire de l'information sur la densité d'échantillonnage.

Soit n points $\mathbb{X}_n = (X_1, \ldots, X_n)$ indépendamment échantillonnés selon une densité f sur $\mathbb{R}^d$. Les complexes de Čech ainsi que les diagrammes de persistance correspondant à la filtration sont des objets aléatoires dont il est naturel d'étudier le comportement asymptotique

lorsque n tend vers l'infini. On se référera à [BK18] pour une étude des comportements asymptotiques de quantités topologiques sur des objets aléatoires. On introduit un terme de renormalisation $(r_n)_{n\in\mathbb{N}}$ qui tend vers 0 afin de rendre la limite non-triviale. La vitesse à laquelle (r_n) tend vers 0 est cruciale : soit $\Lambda := \lim_{n\to\infty} nr_n^d \in [0,\infty]$. Pour notre étude théorique, on va se restreindre au cas $\Lambda = 0$ appelé *régime parcimonieux*, et citer un résultat de [Owa22] :

Théorème II.18. *Soit $\mathbb{X}_n = (X_1, \dots, X_n)$ un échantillon tiré selon une densité g Lipschitz sur $\mathbb{R}^d$, continue et bornée presque partout. Soit $(r_n)_{n\in\mathbb{N}}$ une suite telle que l'on est dans le régime parcimonieux $nr_n^d \to 0$. De plus, on suppose que $n^{k+2}r_n^{d(k+1)} \to \infty$ lorsque $n \to \infty$. Soit $\xi_{k,n}$ le kème diagramme de persistance de $\check{\mathcal{C}}(\frac{1}{r_n}\mathbb{X}_n)$. Soit μ_k la mesure sur $\Delta^+ = \{(x,y) : 0 \le x < y \le \infty\}$ définie sur les rectangles $R_{s,t,u,v} = [s,t)\times[u,v)$ par*

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

pour $0 < s \le t \le u \le v$, où H est une fonction qui dépend de l'arrangement géométrique de ses arguments. On a alors la convergence vague :

$$\frac{\xi_{k,n}}{n^{k+2}r_n^{d(k+1)}} \xrightarrow{v} \mu_k \text{ presque sûrement.}$$

Des résultats similaires sont également présentés dans [Owa22] pour les autres sous-régimes du régime parcimonieux $n^{k+2}r_n^{d(k+1)} \to 0$ et $n^{k+2}r_n^{d(k+1)} \to c > 0$.

Remarque II.19. On peut faire les commentaires suivants sur ce résultat surprenant à première vue :

- La mesure limite μ_k dépend de la densité d'échantillonnage uniquement via la constante multiplicative $\int_{\mathbb{R}^d} f^{k+2}$.
- La constante $\int_{\mathbb{R}^d} f^{k+2}$ est une quantité globale, tandis que les diagrammes de persistance correspondent plutôt à des effects locaux.
- Supposons que l'on cherche à savoir si un nuage de points a été généré selon une densité f_1 ou f_2 telles qu'il existe k tel que $\int_{\mathbb{R}^d} f_1^{k+2} \neq \int_{\mathbb{R}^d} f_2^{k+2}$. En comptant le nombre de points du diagramme de persistance qui tombe dans n'importe quel rectangle (après renormalisation) va bien identifier la densité d'échantillonnage lorsque n est suffisamment grand. En particulier, pour ce problème de classification où l'information homologique est inexistante, le "bruit topologique" contient de l'information discriminante sur l'échantillonnage.

II.3.4 Liste détaillée des contributions

Les contributions de ce travail sont les suivantes :

(i) On a proposé, implémenté et testé une méthode de classification de mesures sans vectorisation. On a comparé sa performance à des méthodes standards d'analyse topologique de données, de classification de séries temporelles, et de cytométrie en flux sur des jeux de données correspondant.

(ii) On a développé une théorie englobant cette méthode et établi des bornes inférieures et supérieures sur la complexité de Rademacher de la classe de fonctions $\tilde{\mathcal{F}}$ définie précédemment.

(iii) On a établi des garanties théoriques propres au problème de classification de diagrammes de persistance, afin de dinstinguer soit des échantillonnages provenant de variétés ayant des groupes d'homologie différents, soit des densités d'échantillonnage différentes sur la même variété. Ce travail a été l'occasion de généraliser le Théorème II.18 à des échantillons sur variétés, tout en proposant une technique de preuve différente de celle du papier original [Owa22].

II.4 Au-delà des diagrammes de persistance : caractéristique d'Euler et multi-persistance

Nous venons de démontrer que les diagrammes contiennent de l'information utile à la fois loin et près de la diagonale, selon les applications visées. Un défaut déjà souligné de la méthode décrite dans la Section II.2 est son temps de calcul. En effet, afin de minimiser la persistance totale Ω_2, il faut calculer un grand nombre de diagrammes de persistance. De plus, on est uniquement intéressé par la quantité $\sum_{(b,d)\in \mathrm{D}}(d-b)$ étant donné un diagramme de persistance D. Ainsi, le calcul précis de toutes les coordonnées de tous les points du diagramme n'est absolument pas nécessaire et beaucoup d'information est volontairement perdue lors du calcul de la persistance totale. Malheureusement, il n'existe pas, à notre connaissance, de façon simple de contourner ce problème et de calculer rapidement la persistance totale sans calculer le diagramme entier. Cette observation a donné lieu à la question suivante :

Question II.20. Quelles statistiques sur les diagrammes de persistance peuvent être calculées en temps linéaire directement des valeurs de filtration et sans calculer le diagramme ? Comment les utiliser pour faire de l'analyse de données ?

II.4.1 Courbes caractéristiques d'Euler et leurs transformées intégrales

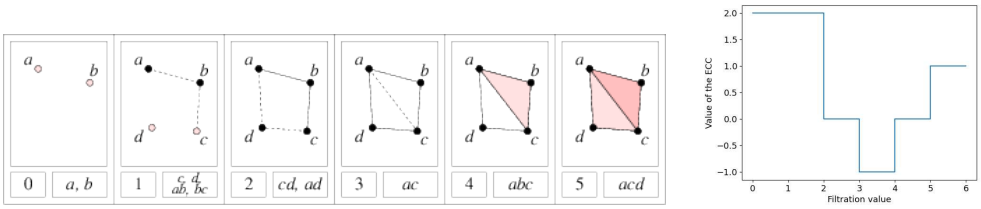
La Question II.20 est au coeur de la Section V, pré-publiée dans [HL23]. Afin de trouver des descripteurs topologiques ne nécessitant pas de calcul de persistance, on revient à la définition de la caractéristique d'Euler d'un complexe simplicial, précédemment défini dans la Définition II.7. On avait remarqué que pour un complexe simplicial $\mathcal{K}$ de dimension d, la quantité $\sum_{k=0}^{d}(-1)^k\beta_k(\mathcal{K})$ peut s'exprimer très simplement comme la somme alternée de tous les simplexes, d'après la formule de la Définition II.7. Cependant, il n'y a pas d'astuce similaire permettant de calculer rapidement la quantité $\sum_{k=0}^{d}\beta_k(\mathcal{K})$, et chaque β_k doit alors être calculé séparément, ce qui résulte en un temps de calcul beaucoup plus grand. Ainsi, il semblerait que le fait d'alterner sur les dimensions permette d'avoir accès à des descripteurs rapides. Dans cette logique, on définit la *Courbe caractéristique d'Euler* ou ECC pour *Euler characteristic curve* :

Définition II.21. Soit $\mathcal{K}$ un complexe simplicial fini muni d'une filtration $(\mathcal{K}_t)_{t\in\mathbb{R}}$. La *courbe caractéristique d'Euler* est la fonction

$$\chi_{\mathcal{K}} : t \in \mathbb{R} \mapsto \chi(\mathcal{K}_t) \in \mathbb{Z}.$$

Sur la Figure 17, on illustre la construction de la courbe caractéristique d'Euler sur une filtration simple issue de [ZC04].

En pratique, la courbe caractéristique d'Euler est vectorisée pour pouvoir être utilisée dans des algorithmes d'apprentissage automatique. Bien que ce descripteur semble assez grossier de prime abord, il a été utilisé plusieurs fois dans la littérature, voir [SZ21, JKN20, AQO+22],



(a) Filtration (b) ECC correspondante

Figure 17: Construction de la courbe caractéristique d'Euler d'une filtration

et nous montrerons dans la Section V que son pouvoir prédictif est essentiellement comparable à celui des diagrammes de persistance, tout en ayant un temps de calcul extrêmement faible.

Afin d'extraire de l'information des courbes caractéristiques d'Euler, on considère des transformées intégrales de ces fonctions, également appelées *transformées hybrides* dans notre contexte, ou HT pour *Hybrid transforms*. Cet objet a été précédemment étudié dans [Leb22] pour quelques unes de ses propriétés théoriques.

Définition II.22. Soit $(\mathcal{K}_t)_{t\in\mathbb{R}}$ une filtration d'un complexe simplicial $\mathcal{K}$ et $\chi_{\mathcal{K}}$ sa courbe caractéristique d'Euler. Soit $\kappa \in L^1(\mathbb{R})$. La *transformée hybride* de noyau κ est la fonction :

$$\psi^{\kappa} : \xi \in \mathbb{R}_+^{\star} \mapsto \xi \cdot \int_{\mathbb{R}} \kappa(\xi s)\chi_{\mathcal{K}}(s)\,\mathrm{d}s.$$

A multiplication par ξ près, les transformées hybrides coïncident avec les transformées intégrales usuelles de courbes caractéristiques d'Euler. Par exemple, si $\kappa = \cos$, cela correspond à la transforme de Fourier en cosinus. Nous montrerons qu'une fois discrétisées, les transformées hybrides sont des descripteurs robustes en analyse de données, plus particulièrement pour faire de l'apprentissage non-supervisé. De plus, le lemme suivant fait le lien entre transformées hybrides et diagrammes de persistance et permet de répondre partiellement à la Question II.20 :

Lemme II.23. *Soit $(\mathcal{K}_t)_{t\in\mathbb{R}}$ une filtration, et ψ^{κ} sa transformée hybride de noyau κ. Soit $\overline{\kappa}$ la primitive de κ telle que $\overline{\kappa}(x) \underset{x\to\infty}{\longrightarrow} 0$. Soit $D_k = \{(b_i^k, d_i^k)\}_{i=1,\dots,n_k}$ le $k$ème diagrame de persistance de $(\mathcal{K}_t)$. Pour un simplexe $\sigma \in (\mathcal{K}_t)$, on note $t(\sigma)$ le premier instant auquel il apparaît dans la filtration. On a alors, pour tout $\xi \in \mathbb{R}_+^{\star}$:*

$$\psi^{\kappa}(\xi) = \sum_{k\geq 0}\sum_{i=1}^{n_k}(-1)^k\Big(\overline{\kappa}(\xi\cdot b_i^k) - \overline{\kappa}(\xi\cdot a_i^k)\Big). \tag{II.1}$$

De plus,

$$\psi^{\kappa}(\xi) = -\sum_{\sigma\in\mathcal{K}}(-1)^{\dim\sigma}\overline{\kappa}\big(\xi\cdot t(\sigma)\big). \tag{II.2}$$

Ce lemme montre que pour tout $\overline{\kappa}$ qui tend vers 0 en l'infini, la fonction définie par l'Equation (II.1) peut être calculée linéairement en les valeurs de filtration par l'Equation (II.2). Si $\overline{\kappa}$ ne tend pas vers 0, on peut toujours utiliser l'Equation (II.2) pour approcher la transformée hybride de noyau κ à constante près. Enfin, en plus d'être des descripteurs rapides et puissants ne nécessitant pas le calcul du diagramme de persistance, les transformées hybrides permettent une compréhension plus profonde du bruit topologique dans les diagrammes de

persistance. En effet, dans la Section II.3, nous avons démontré que les composantes peu persistantes contiennent de l'information liées à l'échantillonnage et à des quantités locales sur la variété. Néanmoins, l'étude de la façon dont ces phénomènes se manifestent dans le bruit topologique est encore ouverte. Les transformées hybrides permettent d'aller un cran plus loin dans l'analyse du bruit topologique, comme nous allons l'illustrer dans l'exemple suivant.

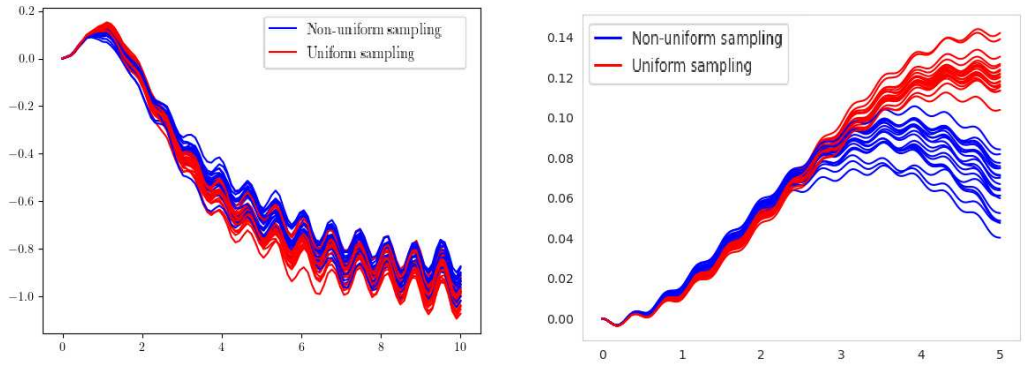
On considère des échantillons de 500 points échantillonnés sur un tore plongé dans $\mathbb{R}^3$. La première classe correspond à des échantillonnages uniformes sur le tore, simulés grâce à un algorithme présenté dans [DHS+13]. La seconde classe correspond à tirer deux angles (θ, φ) uniformément dans $[0, 2\pi]^2$ et obtenir un point sur le tore via l'application $\Psi_{\mathbb{T}^2} : (\theta, \varphi) \mapsto (x_1, x_2, x_3)$, où :

$$\begin{cases} x_1 = (2 + \cos(\theta))\cos(\varphi), \\ x_2 = (2 + \cos(\theta))\sin(\varphi), \\ x_3 = \sin(\theta). \end{cases}$$

Les échantillons de la deuxième classe ne sont pas tirés uniformément sur le tore. On considère un problème similaire sur la sphère, où une classe correspond à 500 points tirés uniformément. Pour l'autre classe, on tire 500 angles θ uniformément dans $[0, \pi]$ et φ selon une loi normale centrée en π. On obtient un point sur la sphère via la paramétrisation sphérique usuelle $\Psi_{\mathbb{S}^2} : (\theta, \varphi) \mapsto (x_1, x_2, x_3)$, où :

$$\begin{cases} x_1 = \sin(\theta)\cos(\varphi), \\ x_2 = \sin(\theta)\sin(\varphi), \\ x_3 = \cos(\theta). \end{cases}$$

Sur la Figure 18, on affiche le début des transformées hybrides de noyau $\kappa = \sin$ pour le complexe de Čech sur quelques échantillons de chaque classe, pour la sphère et le tore.



(a) HT, tore (b) HT, sphère

Figure 18: ECC et HT, deux types d'échantillonnages sur le tore et la sphère

On remarque que les oscillations des transformées hybrides sont en phase et ont la même amplitude pour des données sur la même variété, et apparaissent donc comme des signatures de quantités globales du support de l'échantillonnage. Par ailleurs, sur la même variété, des échantillonnages différents sont décalés verticalement et permettent de distinguer les deux classes. Cette expérience constitue un premier pas vers une compréhension plus profonde des quantités jouant un rôle dans le bruit topologique.

II.4.2 Le défi de la multi-persistence

Jusqu'à présent, nous avons seulement étudié des filtrations de complexes simpliciaux telles que la fonction de filtration est à valeurs réelles, voir la Définition II.8. Cependant, on peut vouloir considérer des sous-niveaux de fonctions multi-valuées définies sur un complexe simplicial. En analyse topologique des données, l'étude de filtrations à plusieurs paramètres est appelée *multi-persistance*. Un des exemples les plus fructueux est de coupler la filtration de Čech avec une fonction sur le nuage de points lui-même, par exemple un estimateur de densité, voir [CB20]. Dans ce cas, l'estimateur de densité permet de filtrer les donnéees aberrantes qui rendent la filtration de Čech instable. Cette extension pourtant simple en apparence pose de nombreux problèmes théoriques et pratiques en persistance

En particulier, d'après [CZ09], il n'existe pas de "diagramme de multi-persistance". Quelques travaux ont adapté des vectorisations usuelles à la bi-persistance, comme les landscapes dans [Vip20], et les images persistantes dans [CB20]. De même, nos outils se généralisent naturellement à la multi-persistance: la courbe caractéristique d'Euler devient le *profil caractéristique d'Euler*, voir [DG22]. Bien qu'il n'y ait désormais plus d'équivalent à l'Equation (II.1), les transformées hybrides peuvent toujours être calculées en utilisant une formule analogue à celle de l'Equation (II.2), et se distinguent désormais des transformées intégrales usuelles. Cette méthode conserve une complexité très compétitive et nous permet d'utiliser jusqu'à cinq filtrations sur certains types de données. En particulier, pour faire de la classification de graphes, de nombreuses fonctions sur les noeuds ou les sommets peuvent mettre en évidence des différences structurelles dans les graphes et peuvent ainsi être utilisées comme filtrations. En plus d'être des vectorisations rapides et puissantes des diagrammes de persistances, les profils caractéristiques d'Euler ainsi que les transformées hybrides deviennent même une nécessité afin de dépasser la contrainte de mono-persistance imposée par les diagrammes de persistance.

II.4.3 Garanties théoriques

Concluons cette partie avec quelques garanties théoriques sur les profils caractéristiques d'Euler et les transformées hybrides. Ces résultats sont de deux natures différentes, et nous allons énoncer ceux liés à la *stabilité* de ces descripteurs, dans le cas-particulier de filtrations par sous-niveaux de fonctions définies sur le même complexe simplicial.

Stabilité Supposons que l'on filtre un complexe simplicial $\mathcal{K}$ via les sous-niveaux de deux fonctions $f, g : \mathcal{K} \to \mathbb{R}^m$, soit χ_f et χ_g les profils caractéristiques d'Euler correspondant et soit ψ_f^κ et ψ_g^κ les transformées hybrides correspondantes, pour un noyau intégrable borné κ. On a les deux résultats de stabilité suivant :

Lemme II.24. *Soit $M > 0$. Alors,*

$$\|(\chi_f - \chi_g)\mathbb{1}_{[-M,M]^m}\|_1 \leq (2M)^{m-1}\|f - g\|_1.$$

De plus, $q \in [1, \infty]$. Il existe une constante C qui dépend seulement de q tel que

$$\|\psi_f^\kappa - \psi_g^\kappa\|_q \leq C\|\kappa\|_\infty\|f - g\|_1.$$

Dans ce lemme, la norme L^1 d'une fonction f définie sur un complexe simplicial $\mathcal{K}$ est définie par $\|f\|_1 = \sum_{\sigma\in\mathcal{K}} \|f(\sigma)\|_1$. Ce lemme montre que ces deux descripteurs sont robustes à des perturbations des filtrations. Malheureusement, ce résultat de stabilité n'est vrai qu'en norme L^1, qui fait intervenir le nombre total de simplexes dans $\mathcal{K}$. Ce résultat de stabilité est ainsi plus faible que la stabilité en norme infinie du Théorème II.15 pour la distance bottleneck.

Théorèmes limites En plus des résultats de stabilité, on montre que les descripteurs de cette section vérifient des garanties asymptotiques lorsque calculés sur des filtrations d'échantillons aléatoires. Dans le cas des courbes caractéristiques d'Euler, ce problème a déjà été étudié dans la littérature. Notamment, dans [KRP21] les auteurs établissent un théorème central limite pour la courbe caractéristique d'Euler. Dans le cas des transformées hybrides, l'Equation II.1 établit que pour un noyau κ et une mono-filtration $(\mathcal{K}_t)$ ayant un diagramme de persistance D_k en dimension $k \in \{0, \dots, d-1\}$, on a

$$\psi^\kappa(\xi) = \sum_{k=0}^{d-1} \langle D_k, h_\xi \rangle,$$

où $h_\xi : (x, y) \mapsto \kappa(\xi y) - \kappa(\xi x)$. Les diagrammes de persistance sont ici traités en tant que mesures discrètes, comme dans la Section II.3. Cette observation élémentaire a les deux conséquences suivantes :

- Les résultats évoqués en Section II.3.2 peuvent être appliqués à l'étude de la capacité des transformées hybrides pour une famille de noyaux.
- Les garanties asymptotiques sur les diagrammes de persistance permettent d'obtenir des résultats limites sur les transformées hybrides sous des hypothèses générales sur le noyau κ. Plus précisément, on peut énoncer le résultat suivant qui combine le Théorème II.18 avec l'observation ci-dessus.

Théorème II.25. *Soit $X_1, \dots, X_n$ un échantillon tiré selon une densité g sur $\mathbb{R}^d$ Lipschitz, continue et bornée presque partout. Soit $(r_n)_{n\in\mathbb{N}}$ une suite telle que $nr_n^d \to 0$ et $n^{k+2} r_n^{d(k+1)} \to \infty$ lorsque $n \to \infty$. Soit ψ_n^κ la transformée hybride de la filtration de Čech de l'échantillon renormalisé $\frac{1}{r_n}(X_i)_{i=1}^n$. Soit $T, a > 0$ et $\kappa \in L^1(\mathbb{R})$. On suppose de plus que le support de κ est inclus dans $[0, T]$. Il existe alors d fonctions $A_0, \dots, A_{d-1}$ dépendant uniquement de κ tel que pour tout $\xi > a$,*

$$\frac{1}{n^{k+2} r_n^{d(k+1)}} \cdot \psi_n^\kappa(\xi) \underset{n\to\infty}{\longrightarrow} \sum_{k=0}^{d-1} \frac{(-1)^k}{(k+2)!} \cdot A_k(\xi) \cdot \int_{\mathbb{R}^d} g^{k+2}(x)\, \mathrm{d}x \quad p.s..$$

Comme pour le Théorème II.18, la densité de l'échantillonnage se retrouve seulement dans les quantités globales $\int_{\mathbb{R}^d} g^{k+2}$ pour $k = 0, 1, \dots, d-1$. Comme pour les diagrammes de persistance, les transformées hybrides permettent de discriminer entre des échantillons issus de densités différentes pour peu que n soit suffisamment grand. Finalement, en Section V, on énoncera un théorème limite en multi-persistance qui découle également de l'observation de l'Equation (II.1).

II.4.4 Liste détaillée des contributions

Nos contributions pour ce travail sont les suivantes :

(i) On a réalisé une étude qualitative approfondie des profils caractéristiques d'Euler et des transformées hybrides sur des données synthétiques et réelles, et étudié l'influence du paramètre de noyau pour les transformées hybrides.

(ii) On a montré que les profils caractéristiques d'Euler permettaient d'obtenir des scores de classification proches de l'état de l'art lorsque mis en entrée d'un classifieur robuste comme une forêt aléatoire tout en ayant un temps de calcul minimal.

(iii) On a montré que les transformées hybrides permettaient de compresser efficacement de l'information, comme le font les transformées de Fourier. En conséquence, elles sont bien plus performantes que les profils caractéristiques d'Euler pour des problèmes non-supervisés ou pour des problèmes supervisés avec une contrainte de budget sur la taille de la vectorisation. Nous avons également illustré leur capacité à extraire de l'information fine sur un jeu de données réel.

(iv) On a établi plusieurs garanties théoriques. On a obtenu des résultats de stabilité similaires à ceux du Théorème II.15 qui explicitent la robustesse de ces objets à une perturbation de la filtration. Nous avons également établi la convergence simple des transformées hybrides associées à des filtrations de Čech sur des échantillons aléatoires ainsi que leur normalité asymptotique. Finalement, nous avons établi une loi des grands nombres pour les transformées hybrides sur des multi-filtrations.

Plan détaillé de la thèse

Cette dissertation est organisée comme suit : dans la Section III, on étudie des problèmes de régression sur variété en utilisant une pénalité topologique. Ce travail est basé sur le paradigme d'une décomposition en signal et bruit topologique discuté en Section II.2. Ce travail a été effectué en collaboration avec Krishnakumar Balasubramanian, Gilles Blanchard, Clément Levrard et Wolfgang Polonik, et a été publié dans [HBB+22]. Dans la Section IV, on a développé une méthode de classification supervisée sur des mesures. On propose deux algorithmes qui s'inscrivent dans un cadre plus générale d'apprentissage statistique sur des mesures, pour lequel on a établi plusieurs garanties théoriques. Ce travail est particulièrement dédié à la classification de diagrammes de persistance, tant pour la théorie que pour les expériences. Il s'agit d'une collaboration avec Gilles Blanchard et Clément Levrard, pré-publié dans [HBL23]. Finalement, dans Section V, on s'affranchit des contraintes computationnelles et de monopersistance des diagrammes en calculant des descripteurs basés sur la caractéristique d'Euler, ainsi que leurs transformées intégrales afin de réaliser diverses tâches d'apprentissage automatique. Ce dernier travail a été réalisé en collaboration avec Vadim Lebovici et a été pré-publié dans [HL23].

III Topologically regularized models on manifolds

In this section, we study how can persistence diagrams help tackle a regression problem on a compact manifold $\mathcal{M}$. In order to take advantage of the underlying geometry and topology of the data, the regression task is performed on the basis of the first several eigenfunctions of the Laplace-Beltrami operator of the manifold, that are regularized with topological penalties. The proposed penalties are based on the topology of the sublevel sets of either the eigenfunctions or the estimated function, as described in Section I.2. The overall approach is shown to yield promising and competitive performance on various applications to both synthetic and real data sets. We also provide theoretical guarantees on the regression function estimates, on both its prediction error and its smoothness (in a topological sense). Taken together, these results support the relevance of our approach in the case where the targeted function is "topologically smooth". The work from this section has been published in [HBB+22] and is joint work with Krishnakumar Balasubramanian, Gilles Blanchard, Clément Levrard and Wolfgang Polonik.

Contents

III.1 Introduction

Problems of regression on manifolds are of growing importance in statistical learning. Given a manifold $\mathcal{M}$, the specific goal is to retrieve a true regression function $f^\star : \mathcal{M} \to \mathbb{R}$ from data X_i (for $i = 1, \dots n$) that lie on the manifold $\mathcal{M}$ and noisy real-valued responses of the form $Y_i = f^\star(X_i) + \varepsilon_i$ where ε_i are the additive noise. Such problems arise in many applications where the data samples X_i, although represented by very high dimensional spaces like sets

of images and 3D volumes, often have an underlying low-dimensional structure and lie on a manifold. This is in particular the case in medical applications. For instance, [GWR+14] and [JZC+15] study regression problems on a set of images of brains. While the set of all images is of very large dimension (the number of pixels), the set of brain images turns out to have a comparatively very small intrinsic dimension. Although there are ways to recover the metric of the underlying unknown manifold [BMTY05], in this section we adopt an extrinsic approach.

A standard approach for estimating $f^\star$ is based on expanding the function in a suitable basis to take advantage of the underlying manifold structure. To this end, we consider the Laplace-Beltrami operator [Ros97], which has been broadly studied, both for its theoretical properties [Zel08, Zel17, SX10] and its great power of applicability in statistical data analysis [Hen90, Lev06, CL06, Sai08, MM07, GBL18, KMK+20]. In our context, we chose the basis to be the set of eigenfunctions of the Laplace-Beltrami operator. Since it is impossible to have access to a closed form expression for the Laplace-Beltrami eigenfunctions for various manifolds in full generality, we replace them by the eigenvectors of the Laplacian matrix of a graph built on the data; see [Moh91] for a complete treatment. Using the eigenvectors of the graph Laplacian for diverse learning tasks is an idea that has its roots in the works of [BN03] and has become extremely popular since. There is a plethora of literature on this topic, and we refer to [WSST15] for a theoretical treatment, and [BNS06] and [CGLS16] for two out of many applications. The use of the graph Laplacian spectrum is backed-up by solid guarantees regarding its convergence towards the spectrum of the Laplace-Beltrami operator. We refer to [vLBB08] for general results adopting the point of view of spectral clustering, [BIK15] for a more recent treatment, and [GTGHS20] for the recent generalization of the latter to random data.

In order to efficiently estimate $f^\star$, we will use a penalization procedure; see [Gir14] or [Mas07] for a complete treatment of these methods. Specifically, we will present two types of penalties that both leverage topological information. These penalties are based on persistent homology, a field that has its origins in algebraic topology and Morse theory [Mil63]. The use of persistent homology has become increasingly popular over the past decade, popularized, among others, by the books [EH22] and [BCY18]. It offers a new approach to data representations. Penalties based on persistence follow a heuristic similar to the one based on total variation (see, for instance, [ROF92, HR16]) which works by reducing the oscillations of the estimated function in order to reconstruct a smooth function. While the heuristics for total variation penalties and persistence based penalties are similar, they still work quite differently, as discussed below.

Penalizing the persistence has been used recently in [CNBW19] for classification applications, and in [BGNDS20] in the context of Generative Adversarial Networks. Furthermore, [CCG+21] has examined optimization with such penalizations in the context of various applications. The novelty of the present work resides in the use of such models in the framework of a regression over a manifold and its joint utilization with a Laplace eigenbasis, enabling a deeper understanding of its topology. It is also the opportunity to study higher dimensional examples where the behavior of topological persistence is fundamentally different from the one of total variation. Indeed, we will see that topological persistence is a very convenient way to prevent the estimated functions from oscillating too much in a stronger way than more standard approaches.

The rest of the section is organized as follows: in an intuitive fashion, Section III.2 presents the motivation behind the introduction of a topological penalty for a Laplace eigenbasis regression and how it can overcome the limitations of total variation denoising. Section III.4

discusses two types of topological penalties: one is equivalent to solving a Lasso problem with weights and therefore has a simple theoretical analysis and even has a closed-form solution, while the other one is non-convex. Despite the lack of guarantees in non-convex optimization, we will present an oracle result for the estimated parameter. We also present a result on controlling the topology, or on topological sparsity, of one of our approaches. In Section III.5, we will present the results of experiments conducted on both synthetic and real data, in order to highlight the strengths and weaknesses of such an approach as opposed to standard regression methods. We have made the code used in several examples available here.[1]

III.2 Motivation

III.2.1 Laplace eigenbasis regression

We study a regression problem on a compact, smooth submanifold $\mathcal{M}$ of dimension d of $\mathbb{R}^D$ without boundary. Throughout this section, we assume the data points $(X_i)_{i=1}^n$ are sampled uniformly and independently over $\mathcal{M}$. Furthermore, for $i = 1, \dots, n$, the responses Y_i are generated based on the model

$$Y_i = f^\star(X_i) + \varepsilon_i, \tag{III.1}$$

where $(\varepsilon_i)_{1 \le i \le n}$ are i.i.d. zero-mean sub-Gaussian noise variables independent of all the X_i's. Our goal is to retrieve the function $f^\star$, also referred to as the regression function, from the given observations $(X_1, Y_1), \dots, (X_n, Y_n)$.

A natural choice of basis to perform a regression and exploit the manifold-structure assumption is the Laplace-Beltrami eigenbasis. Analogously to the Euclidean case, the Laplace-Beltrami operator Δ is (the negative of) the divergence of the gradient: $\Delta f = -\nabla \cdot \nabla f$. If we denote by g the metric tensor and by g^{ij} the components of its inverse, we have the following expression in local coordinates (with Einstein summation convention):

$$\Delta f = -\frac{1}{\sqrt{\det(g)}} \partial_i (\sqrt{\det(g)}\, g^{ij} \partial_j f).$$

We remark here that due to our uniform sampling assumption on the X_i, if suffices to consider the standard Laplace-Beltrami operator as above. The methodology and theory we develop will immediately extend to non-uniform sampling schemes based on *weighted* Laplace-Beltrami operators [Ros97, Gri09] as long as the sampling distribution is sufficiently light-tailed (say, it satisfies Poincaré inequality). In order for our exposition to convey our main contribution on topological penalization, we stick to the uniform sampling assumption in the rest of this section.

Notice that in the Euclidean case, where g is the identity matrix, we retrieve the usual well-known formula for the Laplacian (up to a sign convention). The operator Δ is a self-adjoint operator with compact inverse which implies that its set of eigenvalues is discrete and that they all are non-negative [Ros97]. We can then sort the eigenvalues $(\lambda_j)_{j \ge 1}$ in nondecreasing order and approximate $f^\star$ as a linear combination of the corresponding normalized eigenfunctions $(\Phi_j)_{j \ge 1}$. Besides being an orthonormal basis of $L^2(\mathcal{M})$ with many smoothness properties, it is a known fact that the functions $(\Phi_j)_{j \ge 1}$ are related to the topology of the manifold [Zel08]. In addition, the Laplace-Beltrami eigenbasis can be seen as an extension of the Fourier basis to general manifolds. Indeed, on the two dimensional flat-torus $\mathbb{R}^2/2\pi\mathbb{Z}^2$, the eigenvalues of

[1] https://github.com/OlympioH/Lap_reg_topo_pen

the Laplace-Beltrami operator are $(n^2+m^2)_{m,n\in\mathbb{N}}$ and possible corresponding eigenfunctions are $(x,y)\mapsto \sin(nx)\sin(my)$ up to a normalization constant (see for instance Chapter 4.3 of [Zel17]). By analogy with the approximation of a function by its truncated Fourier series in classical analysis, it is natural to choose the eigenfunctions corresponding to the p smallest eigenvalues as a suitable expansion basis for the signal.

Once the number p of features is chosen, the problem boils down to using the observed data for finding $\theta\in\mathbb{R}^p$ such that $\sum_{i=1}^p \theta_i\Phi_i$ is a good approximation to f^*. To this end, we introduce the design matrix $\mathbf{X}\in\mathbb{R}^{n\times p}$, where $\mathbf{X}_{ij}=\Phi_j(X_i)$, and we let $\hat{\theta}$ be a minimizer of

$$\mathcal{L}(\theta)=\|Y-\mathbf{X}\theta\|_2^2+\mu\Omega(\theta), \tag{III.2}$$

where $Y=(Y_1,\dots,Y_n)$ is the response vector and Ω is a penalty term also depending on the Laplace-Beltrami eigenfunctions. Our choices for Ω will be discussed below. The scalar μ is a calibration factor aiming at reducing overfitting. In case we do not know the eigenfunctions Φ_i, we will use eigenvectors of a graph Laplacian as sample approximation (see below for details).

Examples of classical penalties include L^1-regularization, also called Lasso (see [Bv11] for an exhaustive treatment), and total variation penalty. Although the latter provides good theoretical guarantees (see, for example [HR16] for oracle results and [DN18] for a metric entropy based approach), it fails to capture some aspects of the geometry of the data. Indeed, consider the square $[0,1]^2$ (or equivalently the 2D torus), discretize it as small squares of size ε (we can assume ε to be equal to $1/N$ for some integer N to avoid boundary issues) and consider a pyramidal function f_ε on each square with value 0 at the boundary of the square, and a maximum of ε attained in the middle of the square (see Figure 19). The total variation of the so obtained function is equal to $\sum_{\text{cells}}\int|\nabla f_\varepsilon| = \sharp\text{cells}\int_{\text{cell}}|\nabla f_\varepsilon|$. Since $|\nabla f_\varepsilon|=2$, it yields that $TV(f_\varepsilon)=2$. In particular, it does not depend on ε, which means that total variation is blind to very small perturbations of the function and is therefore not suited to deal with such a type of noise. We are now going to see in the following subsection a type of penalty that can capture such small oscillations.

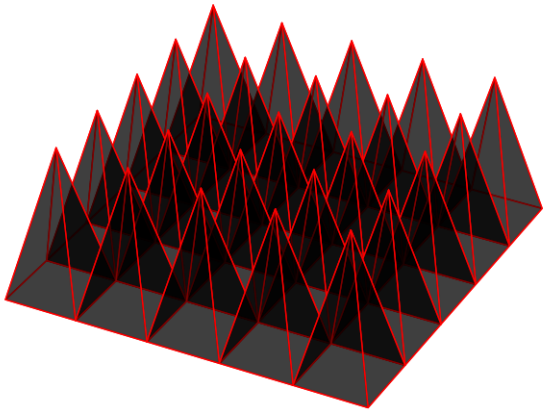


Figure 19: Pyramidal function f_ε.

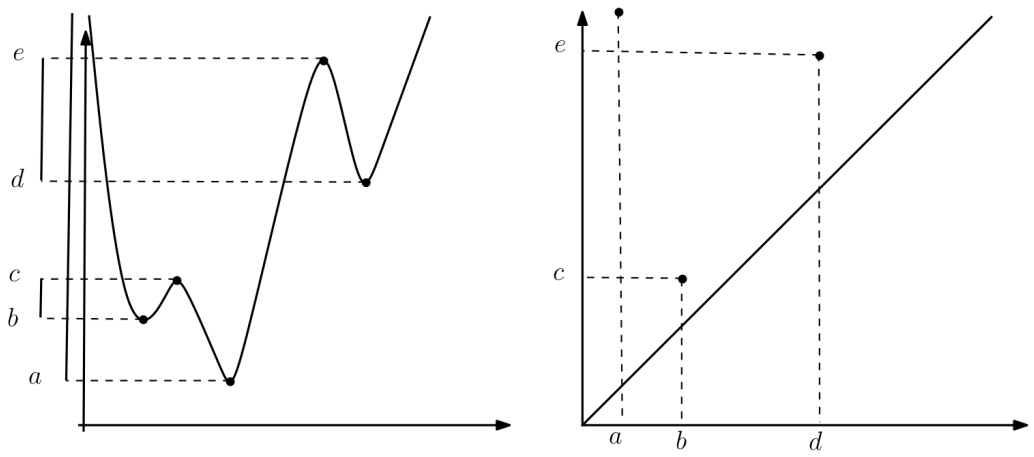


Figure 20: Persistence diagram of a real-valued function.

III.2.2 Total Persistence

In this section, we present the most basic concepts of topological data analysis as introduced in the reference textbooks by [EH22] and [BCY18]. We will try to keep the notions as intuitive as possible and do not lay out the technical details of homology theory. Consider the sublevel sets $\{f^{-1}((-\infty, t])\}_{t=-\infty}^{+\infty}$ of a given function f. As t traverses $\mathbb{R}$ from $-\infty$ to $+\infty$, the topology of the sublevel sets changes and we keep track of these changes in the so-called persistence diagram. More precisely, suppose we are interested in k-dimensional topological features present in a sublevel set (or in a topological space), namely a connected component for $k = 0$, a cycle for $k = 1$, a void for $k = 2$, and so on. For simplicity, assume that f is a Morse function (in particular, its critical points are non-degenerate), such that the topology of the sublevel sets of f only changes at levels t corresponding to extremal points (see Theorem III.1 below). Then, as the level t increases, such k-dimensional features might start to exist at a certain level t_b and they might disappear by merging with another component at a different level t_d, where t_d might be equal to $+\infty$ if it never disappears. Then we place a point in the plane with coordinates (t_b, t_d). The set of all such points (each corresponding to a different k-dimensional feature) along with the diagonal $y = x$ (accounting for the fact that in general features might appear and disappear at the same level or time) forms the $k-$th persistence diagram of f. It is a multi-set of $\mathbb{R}^2$ as different features might appear and disappear at the same levels. The example of a persistence diagram for a one-dimensional function shown in Figure 20 is taken from [BCY18].

For the persistence diagram of a function f to be well-defined, we need the function to satisfy a tameness assumption [EH22]. Sufficient for this is to assume f to be a Morse function. The following result makes precise the above mentioned fact that for Morse functions the topology of the sublevel sets of f can be simply described in terms of its critical points, defined by $\nabla f(x) = 0$. Recall that critical points of Morse functions are non-degenerate (non-singular Hessian), and that the index of a critical point is the number of negative eigenvalues of the Hessian.

Theorem III.1 ([Mil63])**.** *Let f be a Morse function on a smooth manifold $\mathcal{M}$ and denote by $\mathcal{M}^a$ the sublevel set $f^{-1}((-\infty, a])$.*

- *Suppose that there is no critical value between $a < b$. Then $\mathcal{M}^a$ and $\mathcal{M}^b$ are diffeomorphic and $\mathcal{M}^b$ deformation retracts onto $\mathcal{M}^a$.*
- *Suppose p is a non-degenerate critical point of f with index s and that $f(p) = q$. We further assume there are no other critical points p' with $f(p') = q$. Then for ε small enough, $\mathcal{M}^{q+\varepsilon}$ is homotopy equivalent to $\mathcal{M}^{q-\varepsilon}$ with a s-handle attached.*

As a consequence, for Morse functions all the coordinates of the points in the persistence diagrams of every dimension are critical values of f. Furthermore, in the persistence diagram of a feature dimension k, the birth times are critical values of index k and the death times are critical values of index $k+1$. We define the persistence of a feature to be its lifetime, namely its death time t_d minus its birth time t_b, and define the $k-$persistence of a function, denoted by $\mathrm{Pers}_k(f)$, as the sum of all individual persistences in dimension k. In the literature this is sometimes also called the k-total persistence. When we talk about the persistence of a function $\mathrm{Pers}(f)$, it is understood to be the sum of all persistences over all dimensions. The total sum of all the k-th Betti numbers is called the total Betti number of this space.

Note that the existence of features with infinite persistence would make $\mathrm{Pers}(f)$ equal to ∞. To avoid this degeneracy, the quantity $\mathrm{Pers}(f)$ is modified (see [PRSZ20]) by replacing the infinite persistence of a feature born at b by $\max(f) - b$. The number of topological features with infinite persistence equals the total Betti number $\zeta = \zeta(\mathcal{M})$ of the manifold $\mathcal{M}$. In what follows, we will always consider persistences to be clipped as such. We also state a useful result from Chapter 6 of [PRSZ20] in Lemma III.2 below. It can be seen as a corollary of the famous stability inequality in topological data analysis from [CSEH07]. The result essentially states that two functions close in uniform norm necessarily have close persistence.

Lemma III.2 ([PRSZ20])**.** *Let f and h be two Morse functions on a manifold $\mathcal{M}$ with total Betti number ζ. Denote by $\nu(f)$ the total number of points (with finite persistence) in the persistence diagram of f. Then*

$$\mathrm{Pers}(f) - \mathrm{Pers}(h) \leq (2\nu(f) + \zeta)\|f - h\|_\infty.$$

This result remains true when f is Morse and h is only continuous. Under those circumstances, $\mathrm{Pers}(h)$ can be defined by the (possibly infinite) limit of the total persistence of a sequence of Morse functions that uniformly converges towards h, as done in [PPS19]. It is worth mentioning here that more precise stability results for difference of total persistences with respect to the L_1 metric (instead of L_∞) are available in [ST20], in the case where functions are defined on top of CW-complexes. Though adaptation of such results to sublevel sets based filtration seems possible, applications to this particular case of regression on manifold would lead to the same kind of bounds. Indeed, Lemma III.8 ensures that sup-norm bounds are of the same order as L_1 bounds in this case. Nonetheless, we believe that substantial gains might be expected from using these refined bounds in more general regression settings.

III.3 Methodology

In applications we construct persistence diagrams from random data — think of a random function, such as an estimated regression function, or a function with noise added; see below. The standard paradigm in topological data analysis is that in such random persistence diagrams the features with a high persistence are true features, whereas the features with a low persistence that lie near the diagonal are noisy perturbation of the topology. We denote that recent results from [BHPW20] are changing this paradigm since relevant topological information can be found in low-persistence features. Though it is likely that some local information

may be retrieved from these small persistence features (such as geometrical characteristics of the support), in a regression setting given a noisy input, topological smoothness of the regression function is enforced via discarding these small oscillations. We propose two penalization strategies that intend to achieve this goal. We can see an example of the influence of noise on the persistence diagram Figure 21 where we have computed the value of a function at 1000 points uniformly sampled in the square $[0, 10]$ and where we have added Gaussian noise to each entry with three different levels ($\sigma = 0.01, \sigma = 0.05$, and $\sigma = 0.1$), and then plotted an interpolation. The function considered here is the sum of four Gaussian functions on a square. This function has a single topological feature of dimension 0 (a connected component is born at level 0 and never dies) and four topological components of dimension 1 (that die at the height of the local modes of each Gaussian). When adding noise to this function, the resulting persistence diagram has many points and the noisy function has a very large persistence. The higher the noise level, the further the noisy features are from the diagonal, until it is hard to distinguish the four true topological features from the noisy features. We can see this observation reflected in the plots of the function itself. This motivates to consider methods that sparsify the persistence diagram in order to denoise the input.

III.3.1 Two types of penalties

We will introduce two different ways of penalizing the persistence. The first one aims at reducing the dimension of the problem by selecting a 'small' number of eigenfunctions, while the second one is more focused on denoising and providing a smoother output.

We first consider the penalty

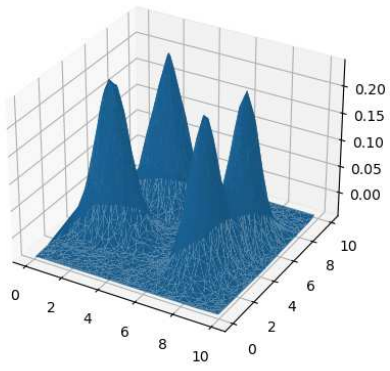
$$\Omega_1(\theta) = \sum_{i=1}^{p} |\theta_i| \mathrm{Pers}(\Phi_i), \tag{III.3}$$

where the $(\Phi_i)_{i=1}^p$ are eigenfunctions of the Laplace-Beltrami operator. From a theoretical viewpoint, every homological dimension should be penalized in order to capture every possible oscillation of the regression function. To give an illustration, considering the pyramidal oscillations of the function depicted in Figure 23 upward oscillations are captured by homology of dimension one, whereas downward ones may be seen on the 0-dimensional persistence diagram. Depending on the problem at hand, the persistence of only one or a few chosen homological dimensions can be penalized as we will see in Section III.5. When treating high-dimensional data, it actually becomes a computational necessity to only penalize by the first homological dimensions, as we will discuss in Section III.5.4. The idea behind the penalty is that the more a function oscillates, the more likely it is to overfit the data. The penalty Ω_1 can be understood as a weighted Lasso penalty, with weights being the persistences of each eigenfunction. The weighted Lasso is a broadly studied model (see [Bv11] for an exhaustive reference). It induces sparsity in the representation of the function, and in our context it aims at introducing an inductive bias towards discarding eigenfunctions with a large persistence.

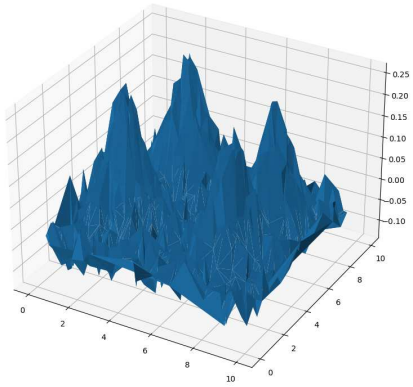
The second persistence based penalty considered here is

$$\Omega_2(\theta) = \mathrm{Pers}\left(\sum_{i=1}^{p} \theta_i \Phi_i\right). \tag{III.4}$$

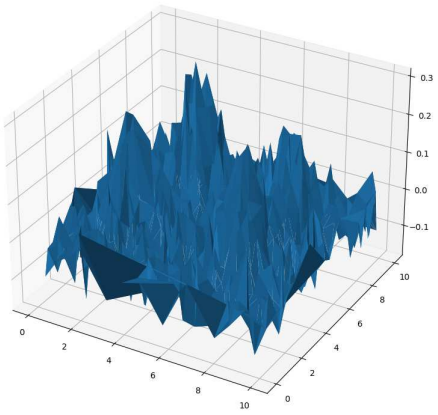
While Ω_2 does not induce sparsity over the parameter θ, it aims at inducing a certain kind of 'topological sparsity'. Indeed, the goal of this penalty is for the reconstructed function to have a much smaller number of points in the persistence diagram than the noisy function.



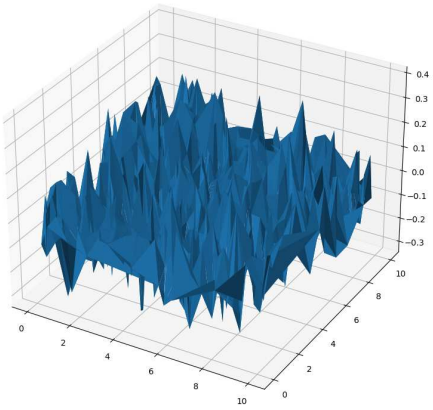
(a) Original function



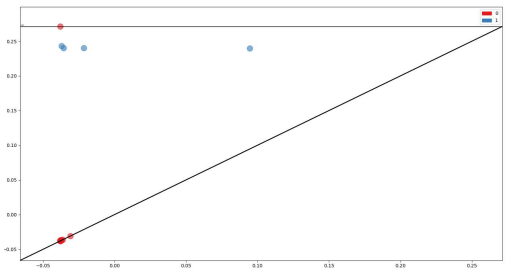
(b) $\sigma = 0.03$



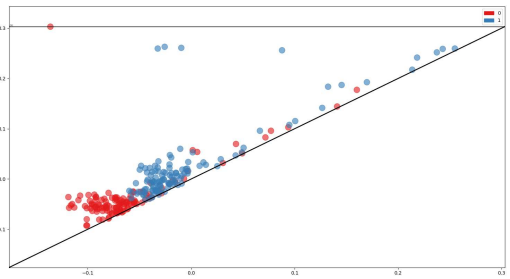
(c) $\sigma = 0.05$



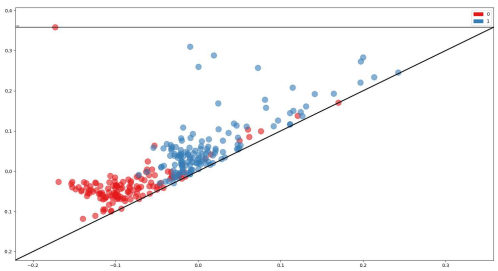
(d) $\sigma = 0.1$



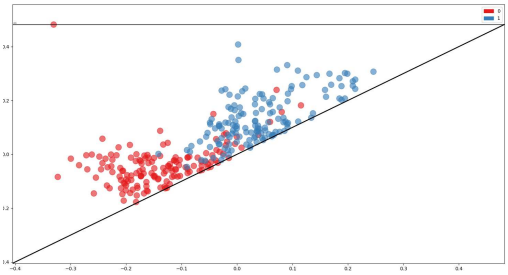
(e) Original function



(f) $\sigma = 0.03$

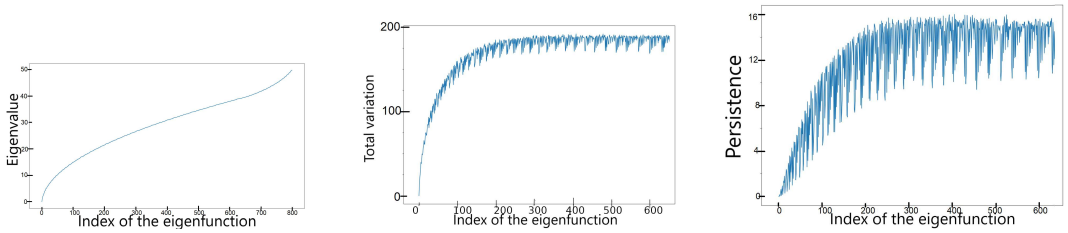


(g) $\sigma = 0.05$



(h) $\sigma = 0.1$

Figure 21: Influence of noise on persistence diagrams.



(a) Eigenvalue (b) Total variation (c) Persistence

Figure 22: Various quantifiers of the oscillations of the eigenfunctions of the Laplacian.

Intuitively, this is achieved by optimizing θ so as to cancel out oscillations of the individual eigenfunctions Φ_i when summing them up. A main computational issue with the penalty Ω_2 is its non-convexity. Indeed, if we consider the two functions $\cos(nx)$ and $\cos(ny)$, they both have a persistence of order n, yet the sum of the two has a persistence of order n^2. However, a result in [CCG+21] shows the convergence of a stochastic gradient descent algorithm towards a critical point of the loss function for the penalty Ω_2.

Finally, we remark that one can potentially consider variants of penalty Ω_1, for instance, a weighted Ridge penalty of the form $\sum_{i=1}^{p} \theta_i^2 \mathrm{Pers}(\Phi_i)$. Preliminary experiments have shown that the weighted Lasso performs better than the weighted Ridge. In addition, as we will see in Section III.5, in applications a combination of the two penalties described above is quite efficient, and we actually benefit of the selection properties of the weighted Lasso.

III.3.2 Contrasting persistence and total variation

While persistence at a first glance might appear to be a measure of the regularity of a function similar to total variation, this only is true for a function in one dimension, where the persistence is half the total variation for functions on the circle $\mathbb{S}^1$ (see [PRSZ20]). In higher dimensions these two penalties are no longer equivalent as discussed in the following.

A first indication of the differences between total variation and persistence is given in Figure 22, showing Laplace-Beltrami eigenvalues on the flat torus along with persistences and total variations of their eigenfunctions $\sin(nx)\sin(my)$. Note that for this figure, the persistence and the total variation have been computed numerically for eigenfunctions defined on a regular grid. Within an eigenspace, the eigenvalues are sorted in lexicographical order on (n, m). We defer to Section III.5.1 for more details on how to numerically compute persistences. We remark here that the x-axis (corresponding to the index of eigenfunctions) in the sub-figures are all aligned.

Figure 22 shows that the eigenvalues increase 'smoothly' and using them as weights for a Lasso-type penalty is a way to regularize the oscillatory behavior of eigenfunctions of large index. However, it can be seen in panel (c) in Figure 22 that while the persistences of the eigenfunctions show an increasing trend with increasing eigenvalues, the persistences also show an overlaid periodic behavior. A similar behavior can be seen for the total variation, but with a much smaller periodic effect. The significant periodic behavior of the persistences means that eigenfunctions can have similar persistences, even if their eigenvalues are quite different. Vice versa, eigenfunctions with similar (or equal) eigenvalues can have quite different persistences. Indeed, for even fixed integers n and m, let $\Phi : (x, y) \mapsto \sin(nx)\sin(my)$ be a corresponding eigenfunction. Its gradient is $\nabla\Phi(x, y) = (n\cos(nx)\sin(my), m\sin(nx)\cos(my))$

and it therefore has $2nm$ critical points. By using the fact that the number of saddle points must be equal to the sum of the number of maxima and the number of minima because the Euler characteristic of the torus equals 0, we obtain that Φ has exactly $nm/2$ maxima, $nm/2$ minima and nm saddles. One of the minima, two of the saddle points, and one of the maxima generate essential homology classes whose corresponding persistent homology classes live forever. Following the convention taken, those are truncated at the maximum value of the function. The persistence diagram of dimension 0 has a point of persistence 2 and $mn/2-1$ of persistence 1. Therefore, the 0-persistence is $mn/2+1$. For homological dimension 1, we have $mn/2+1$ points, all of them have persistence 1, so the 1-persistence is $mn/2+1$. For homological dimension 2, the only point in the persistence diagram has coordinates $(1,1)$. The total persistence is therefore $mn+2$. The case where n or m is odd is very similar and also yields that the persistence is of order mn. This means that within an eigenspace with eigenvalue $\lambda = n^2+m^2$, a penalty on the persistence is proportional to mn, therefore eigenfunctions $\sin(nx)\sin(my)$ with n or m small are more likely to be kept in the model. For instance, the eigenfunctions $(x,y) \mapsto \sin(10x)\sin(y)$ and $(x,y) \mapsto \sin(8x)\sin(6y)$ correspond to eigenvalues 101 and 100 but have very different persistence, namely five times larger for the latter eigenfunction. This effect is much less pronounced for the total variation penalty.

While the above already shows some differences between the two types of measures of regularity of functions, the following observation is perhaps even more relevant for our purposes. To this end, let us reconsider the example of the 2-dimensional pyramidal function f_ε shown in Fig. 19. We already observed above that the TV-penalty does not depend on the choice of ϵ. To understand the behavior of the persistences of these functions, observe that the sublevel sets of the function f_ε are empty for levels $t<0$, and then for $t \in (0,\varepsilon)$, the sublevel sets have $1/\varepsilon^2$ homology components of dimension 1, that all merge at ε. Therefore, the 1-persistence diagram of f_ε has $1/\varepsilon^2$ points, all born at level 0 and dying at time ε. Thus, f_ε has a 1-persistence of $1/\varepsilon$, which thus increase to infinity as $\epsilon \to 0$. This is in stark contrast to the behavior of the total variation. In a similar fashion, the sublevel sets of the function $-f_\varepsilon$ have $1/\varepsilon^2$ connected components from $-\varepsilon$ to 0 that all merge at 0. Therefore, $\mathrm{Pers}_0(-f_\varepsilon) = 1/\varepsilon$ while it also has a total variation that does not depend on ε.

As an example of a function where both Pers_0 and Pers_1 are of importance, consider the same discretization of the space as above, where this time we alternate between a pyramid of height ε and a reversed pyramid of height $-\varepsilon$ (see Figure 23). Similarly to the two previous cases, the persistence diagram of this function has $1/(2\varepsilon^2)$ points of coordinate $(-\varepsilon,\ 0)$ (for the 0-homology) and $1/(2\varepsilon^2)$ points of coordinate $(0,\ \varepsilon)$ (for the 1-homology). Therefore, its 0-persistence is equal to its 1-persistence, both equal to $1/(2\varepsilon)$, while the total variation here again does not vary with ε.

$\|\cdot\|_\infty$	Lip	TV	Pers_1
ε	2	2	$1/\varepsilon$

Table 1: Quantities characterizing f_ε.

It is straightforward to build similar examples in higher dimensions and the effect will be even more striking: For instance, discretize the $d-$dimensional hypercube with cubes of size $1/\varepsilon^d$ and consider a function increasing linearly towards the center of each cube until it reaches a maximum of ε. Such a function still has a total variation of order 1 no matter the value of ε, however, its $(d-1)$-persistence will be equal to $1/\varepsilon^{d-1}$.

The takeaway of the above discussion is as follows: Consider Table 1 which provides a summary of various measures of regularity of the pyramidal function f_ε. We see that if we penalize the supremum norm of this function, the penalty will have no effect as $\varepsilon \to 0$. If we try to penalize its Lipschitz constant or its total variation, the penalty will be the same,

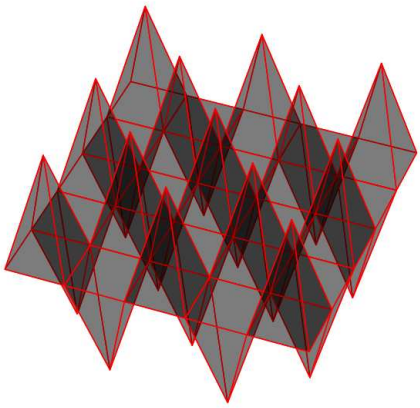


Figure 23: Alternating pyramids.

no matter the scaling ε and it will therefore have a very limited effect. In contrast to that, when penalizing the persistence, the effect of the penalty will become quite important as ε becomes small. Such a function f_ε is assimilated to noise as $\varepsilon \to 0$ and we want to penalize it as much as possible, which is something that can be achieved by persistence but not by total variation.

III.3.3 Complexity of functions with bounded persistence: A negative result

When penalizing persistence, a natural question that immediately arises is to measure the *complexity* of the set of bounded-persistence functions. Loosely speaking, if the set of candidate functions has a large complexity, seeking for a candidate function (e.g. minimizing an empirical loss) can be very challenging and furthermore, the control of the excess risk between $f^\star$ and its estimation becomes non-informative (see [MRT18, Sections 3 and 11] for a more detailed exposition). For regression problems, a standard measure of the size (complexity) of a set of functions $\mathcal{F}$ is the so-called fat-shattering dimension introduced in [KS94].

Definition III.3. Let $\gamma > 0$. A set of points $\mathbb{X} = \{X_1, \ldots, X_l\}$ is said to be γ-shattered if there exists thresholds $r_1, \ldots, r_l$ such that for any subset $E \subset \mathbb{X}$, there exists a function $f_E \in \mathcal{F}$ such that $f_E(x_i) \geq r_i + \gamma$ if $x_i \in E$ and $f_E(x_i) < r_i - \gamma$ if $x_i \notin E$ for all i. The fat-shattering dimension $\mathrm{fat}_\gamma(\mathcal{F})$ of the class $\mathcal{F}$ is then equal to the cardinality of the maximal γ-shattered set X.

Note that the fat-shattering dimension of the class $\mathcal{F}$ depends on the parameter $\gamma > 0$. A class $\mathcal{F}$ has infinite fat-shattering dimension if there are γ-shattered sets of arbitrarily large size. It is well-known that bounds on the fat-shattering dimension lead to bounds on the covering number and hence the metric entropy and Rademacher complexity of the function class. Furthermore, [BLW96] showed that a function class $\mathcal{F}$ is learnable (in the sense of [BLW96, Definition 2]) if and only if it has finite fat-shattering dimension. Unfortunately, in the case of bounded persistence functions, we have the following result:

Theorem III.4. *Let* $\mathcal{H}_V = \{f : [0,1]^d \to [0,1] | \, \mathrm{Pers}(f) \leq V\}$. *Let* $0 < \gamma < 1/2$.

- *If* $d = 1$, $\quad \mathrm{fat}_\gamma(\mathcal{H}_V) \leq 1 + \lfloor \frac{V}{\gamma} \rfloor$.
- *If* $d \geq 2$, *then* $\mathrm{fat}_\gamma(\mathcal{H}_V) = \infty$ *if* $2\gamma \leq V$, *and* $\mathrm{fat}_\gamma(\mathcal{H}_V) = 1$ *otherwise.*

Proof. First we note that if $d = 1$, for any function $f : [0,1] \to [0,1]$, we have that $\text{Pers}(f) = \frac{1}{2}(TV(f) + |f(1) - f(0)|)$, and therefore:

$$\frac{1}{2}TV(f) \leq \text{Pers}(f) \leq TV(f).$$

We can therefore derive the claim for $d = 1$ using the fact that the $\gamma-$fat shattering dimension of the set of functions with total variation smaller than V is equal to $1 + \lfloor V/2\gamma \rfloor$. We refer to Corollary 4.3 from [Sim97] for a detailed proof. Note that if we had considered functions on the circle, by effectively setting $f(0) = f(1)$, we would have had $\text{Pers}(\cdot) = \frac{1}{2}TV(\cdot)$, and therefore the claim for $d = 1$ would be an equality.

In general, in the case where $2\gamma > V$, a point can be shattered by constant functions; on the other hand, given two points x, y that are shattered, there must exist two real numbers r_x, r_y and two functions f, g in the family such that $f(x) > r_x + \gamma > r_x - \gamma > g(x)$ and $f(y) < r_y - \gamma < r_y + \gamma < g(y)$. Thus (depending if $r_x \geq r_y$ holds, or the opposite) either f or g must necessarily have a range of values larger than 2γ and therefore its persistence must be larger than 2γ, which yields a contradiction. Hence, in any dimension $\text{fat}_\gamma(\mathcal{H}_V) = 1$ if $2\gamma > V$.

Assume now $d = 2$ and $2\gamma \leq V$. Consider a set of n points $x_1, \ldots, x_n$ in $[0,1]^2$ that form a regular $n-$gon. Let $E \subseteq \{1, \ldots, n\}$ be an arbitrary subset of indices. We consider a function f such that

$$\begin{cases} f(x) = -V/2 \text{ if } x \in \text{Conv}(x_i)_{i \notin E}, \\ f(x_i) = V/2 \text{ if } i \in E, \end{cases}$$

and f increases smoothly on $\text{Conv}(x_i)_{i=1}^n \setminus \text{Conv}(x_i)_{i \notin E}$ and if $x \notin \text{Conv}(x_i)_{i=1}^n$, $f(x) = f(\Pi_{\text{Conv}(x_i)_{i=1}^n}(x))$ where $\Pi_{\mathcal{C}}(x)$ denotes the projection of x onto a convex set $\mathcal{C}$. The function f defined as such has a persistence of V and the set $\mathcal{H}_V$ therefore γ-shatters this set of n points. Similar examples can be built for $d > 2$. □

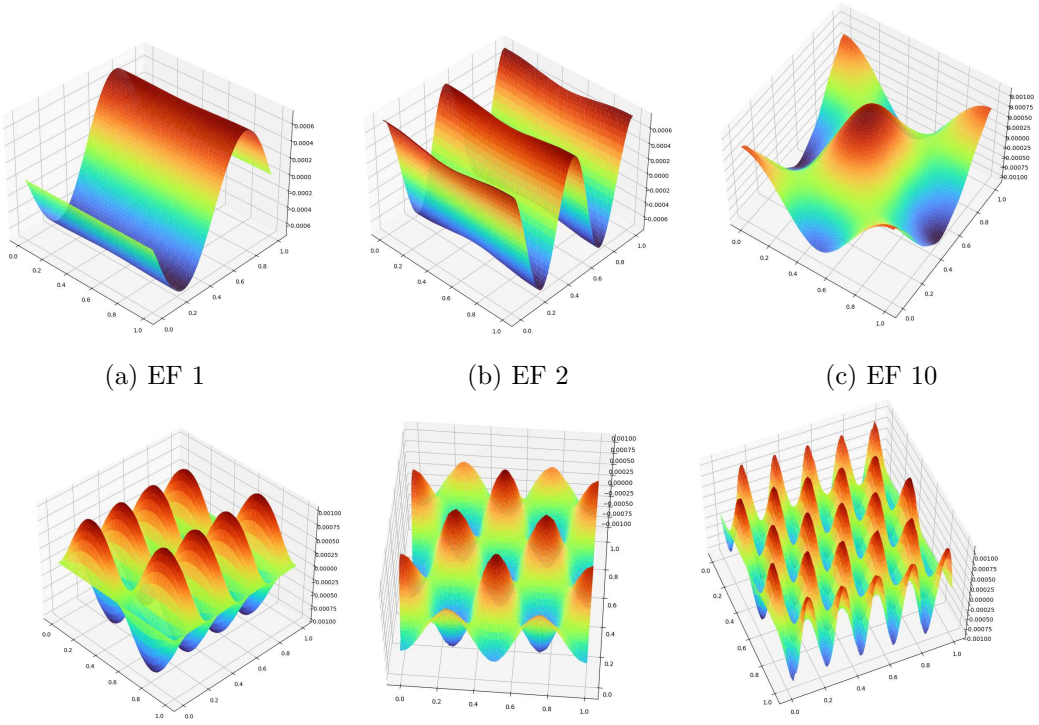
This observation highlights a challenge to overcome when constructing penalties involving topological persistence. This serves as our main motivation for our proposed penalties, in order to restrict the size of the set of candidate functions based on eigenbasis expansions.

III.3.4 Empirical eigenfunctions

For simple manifolds (a flat open space, a torus or a sphere for instance), computing the spectrum of the Laplace-Beltrami operator is analytically tractable. However, for general manifolds this is not possible. Moreover, in practical problems, the manifold itself may be unknown. To deal with this, we take the standard empirical approach and build an undirected graph on the vertex set $V = \{X_1, \ldots, X_n\}$, with weights W_{ij} between the vertex i and the vertex j that are computed according to the ambient metric. In the following experimental study, we consider nearest neighbor and Gaussian-similarity based graphs. We denote by $L = D - W$ the unnormalized graph Laplacian matrix with degree matrix D and weight matrix W. The degree matrix is a diagonal matrix simply defined as

$$D_{ii} = \sum_{j=1}^{n} W_{ij} \text{ and } D_{ij} = 0 \text{ if } i \neq j.$$

The normalized Laplacian matrix is then given by $L' = D^{-1/2}LD^{-1/2}$. Both the matrix L and L' are symmetric positive semi-definite and therefore admit a basis of orthogonal



(a) EF 1 (b) EF 2 (c) EF 10

(d) EF 27 (e) EF 54 (f) EF 75

Figure 24: Several estimated eigenfunctions of the graph Laplacian, 10000 points sampled on $\mathbb{T}^2$, 8-NN graph.

eigenvectors. We will only focus on the normalized Laplacian since it provides slightly better convergence guarantees [vLBB08]. The use of these eigenvectors is justified by the fact that they converge to the true eigenfunctions of the Laplace-Beltrami operator in various metrics as the number of points n tends to infinity and the scaling parameter of the graph tends to 0 [Kol98, GTGHS20]. A few estimated eigenfunctions based on nearest-neighbor graph Laplacian are plotted in Figure 24, for points regularly sampled on the unit square folded into a torus. The nearest-neighbor graph is built thanks to the ambient metric of $\mathbb{R}^3$ (and not the metric on the torus). This is justified because the results of [GTGHS20] ensure the convergence of the spectrum of the graph Laplacian built on the ambient metric. This is due to the fact that locally, the metric on the manifold resembles the ambient metric (see Proposition 2 of [GTGHS20]). In what follows, the $i-$th eigenvector of the Graph Laplacian matrix is denoted by $\hat{\Phi}_i$.

We also remark that while we previously defined the persistence for the true Laplace-Beltrami eigenfunctions, we can simply extend the definition for estimated graph-Laplacian eigenfunctions on V by considering $\overline{f}_i := \sum_{j=1}^{n} \hat{\Phi}_i(X_j)\mathbb{1}_{V_j}$ where V_j is the Voronoi cell centered on X_j. We also use the notation $\mathrm{Pers}(\hat{\Phi}_i) = \mathrm{Pers}(\overline{f}_i)$. Finally, we also mention that using the spectrum of the Laplacian of a graph is a broadly developed idea to perform various statistical learning tasks such as regression or clustering; see, for example, [CGLS16], [IS14], [NJW02] and [vLBB08].

III.4 Theoretical guarantees

We now provide novel theoretical guarantees for the proposed persistence regularization methodology.

Assumption 1. *We assume the true regression satisfies one of the two assumptions below.*

A1 *There exists* $\theta^\star \in \mathbb{R}^p$ *with* $\|\theta^\star\|_0 = s$ *such that* $f^\star = \sum_{i=1}^p \theta_i^\star \Phi_i$*, where* Φ_j *are the eigenfunctions of the Laplace-Beltrami operator with corresponding eigenvalue* λ_j*. In this case, we say that* $f^\star$ *has a sparsity index of* s *over the basis* $(\Phi_i)_{i=1}^p$*.*

A2 *There exists* $\theta^* \in \mathbb{R}^p$ *such that* $f^\star = \sum_{j=1}^p \theta_j^\star \Phi_j$ *and* $f^\star$ *is a Morse function.*

A remark is in order regarding the sparsity assumption on f^* in Assumption 1-**A1**. As discussed in Section III.3.2, the relationship between the topological regularity of the eigenfunctions and their ordering is not known to be monotone in general, and it only exhibits a periodic trend. Hence, to maintain generality, we assume f^* is a sparse linear combination of the eigenfunctions.

III.4.1 Theoretical guarantees for the Ω_1 penalization

We are first interested in the properties of the penalty Ω_1 introduced in Section III.3.1. This approach can simply be understood as a weighted Lasso with a random design. Since the Laplace-Beltrami eigenfunctions form a basis of $L^2(\mathcal{M})$, the compatibility condition [vdGB09] is verified and we have a fast rate of convergence.

Theorem III.5. *Assume that* f^* *satisfies Assumption 1-****A1****. Assume we observe* $Y_i = f^\star(X_i) + \varepsilon_i$ *where* ε_i *are zero-mean sub-Gaussian random variables with parameter* σ^2*. Let* $\hat{\theta}$ *be the minimizer of* $\mathcal{L}$ *given by (III.2) with penalty* Ω_1 *given by (III.3). Then there exists a constant* $C(\mathcal{M})$ *that depends only on the manifold* $\mathcal{M}$ *such that for all* $x > 0$*, we have that if*

$$\mu \geq 2\sigma \sqrt{\frac{pC(\mathcal{M})(\ln(p) + x)}{n}},$$

then with probability larger than

$$1 - 2e^{-x} - e^{-\frac{0.15n}{C(\mathcal{M})p} + \ln(p)},$$

we have

$$\frac{1}{n}\|\mathbf{X}(\hat{\theta} - \theta^\star)\|_2^2 + \mu \sum_{i=1}^p \mathrm{Pers}(\Phi_i)|\hat{\theta}_i - \theta_i^\star| \leq \mu^2 \frac{s}{2}.$$

The proof of the above theorem is provided in Section III.6. For the result in Theorem III.5 to hold with large probability, the number of samples n should be of order at least $p\ln(p)$. Under those circumstances, if the trade-off parameter μ is chosen of order $\sigma\sqrt{\frac{p\ln(p)}{n}}$ as this theorem suggests, it can be shown that the overall prediction error is of order $\frac{\sigma^2 p \ln(p) s}{n}$, up to multiplicative constants [vdGB09]. According to Lemma III.8 in Section III.6, the use of a Laplace-Beltrami eigenbasis here translates into an additional multiplicative factor of p as opposed to a standard Lasso with a design matrix satisfying the RIP conditions [Bv11].

In the case where we study an approximation of the Laplace-Beltrami eigenbasis by using the estimated eigenfunctions $\hat{\Phi}_i$ of the graph Laplacian, the design matrix can be chosen to

be orthonormal. Under those circumstances, the Lasso has an explicit solution, which we illustrate next. Let a design matrix $\hat{\mathbf{X}}$ be built on the estimated eigenfunctions or the graph Laplacian eigenvectors. Then, the minimizer $\hat{\theta}$ of the functional

$$\mathcal{L}'(\theta) = \|Y - \hat{\mathbf{X}}\theta\|_2^2 + \mu \sum_{i=1}^{p} |\theta_i| \text{Pers}(\hat{\Phi}_i), \tag{III.5}$$

if a soft-thresholding type estimator given for all j by:

$$\hat{\theta}_j = \text{Pers}(\hat{\Phi}_j)\hat{\Phi}_j^T Y \left(1 - \frac{\mu \text{Pers}(\hat{\Phi}_j)}{2|\hat{\Phi}_j^T Y|}\right)_+. \tag{III.6}$$

To see this, consider a new design matrix $\hat{\mathbf{X}}$ such that $\hat{\mathbf{X}}_{i,j} = \hat{\Phi}_j(X_i)/\text{Pers}(\hat{\Phi}_j)$. Minimizing the functional in (III.5) is then equivalent to minimizing the functional

$$\mathcal{L}'(\theta) = \|Y - \hat{\mathbf{X}}\theta\|^2 + \mu\|\theta\|_1,$$

by solving a standard Lasso problem. This function is convex in θ and therefore admits a global minimum (not necessarily unique) that we will denote by $\hat{\theta}$. Although $\mathcal{L}'$ is not differentiable because of the L^1 penalty, we can still write the optimality conditions in terms of its sub-differential. Specifically, we have the sub-differential to be

$$\partial\mathcal{L}'(\theta) = \{-2\hat{\mathbf{X}}^T(Y - \hat{\mathbf{X}}\theta) + \mu z : z \in \partial\|\theta\|_1\},$$

from which the optimality condition could be obtained as

$$\hat{\mathbf{X}}^T\hat{\mathbf{X}}\hat{\theta} = \hat{\mathbf{X}}^T Y - \frac{\mu}{2}\hat{z},$$

where $\hat{z} \in \mathbb{R}^p$ is such that $\hat{z}_j = \text{sgn}(\hat{\theta}_j)$ whenever $\hat{\theta}_j \neq 0$ and $\hat{z}_j \in [-1, 1]$ whenever $\hat{\theta}_j = 0$. Due to the orthogonal design, the term $\hat{\mathbf{X}}^T\hat{\mathbf{X}}$ simplifies and a straightforward analysis of the possible cases given the sign or the nullity of $\hat{\theta}_j$, for all j leads to the solution of θ_j given in (III.6) for all j. We therefore have an explicit condition on the eigenbasis selection process, that is $\hat{\theta}_j = 0$ if and only if $|\langle\hat{\Phi}_j, Y\rangle| \leq \frac{\mu}{2}\text{Pers}(\hat{\Phi}_j)$. Stated otherwise, an eigenvector with a high persistence has to explain the data significantly well to be kept in the model.

III.4.2 Theoretical guarantees for the Ω_2 penalization

The penalty Ω_2 being non-convex, a thorough theoretical study seems more complicated. Nonetheless guarantees on the prediction error and on the persistence of the reconstruction can be derived, as described below.

Theorem III.6. *Let f^* satisfy Assumption 1-**A2**. Assume we observe $Y_i = f^\star(X_i) + \varepsilon_i$ where ε_i are zero-mean sub-Gaussian random variables with parameter σ^2. We further assume that for all $i = 1, \dots, n$, X_i is sampled uniformly from $\mathcal{M}$ and that ε_i is independent of X_i. Let $\hat{\theta}$ be the minimizer of $\mathcal{L}$ given by (III.2) with penalty Ω_2 given by (III.4). Then for all $x > 0$, the estimated parameter $\hat{\theta}$ verifies with probability larger than*

$$1 - 2e^{-x} - \exp\left(\frac{-0.1n}{C(\mathcal{M})p} + \ln(2p)\right),$$

that

$$\|\theta^\star - \hat{\theta}\|^2 \leq 16\frac{p\sigma^2}{n}\left[1 + C(\mathcal{M})\sqrt{\frac{2x}{n}}\right](1+\sqrt{x})^2 + 4C(\mathcal{M})p\,(2\nu(f^\star)+\zeta)^2\mu^2.$$

Here, we recall that μ is the trade-off parameter, $C(\mathcal{M})$ is a constant that depends only on the manifold $\mathcal{M}$, ζ is the total Betti number of $\mathcal{M}$ and $\nu(f^\star)$ is the number of points in the persistence diagram of $f^\star$. In addition, we also have under the same hypotheses with $\hat{f} = \sum_{j=1}^{p}\hat{\theta}_i\Phi_i$:

$$\mathrm{Pers}(\hat{f}) \leq \mathrm{Pers}(f^\star) + 16\frac{p\sigma^2}{\mu n}\left[1 + C(\mathcal{M})\sqrt{\frac{2x}{n}}\right](1+\sqrt{x})^2 + 8C(\mathcal{M})p(2\nu(f^\star)+\zeta)^2\mu.$$

The proof of the above theorem is provided in Section III.6. This result holds with large probability if the number of samples n is at least of order $O(p\ln(p))$. Choosing $\mu = O(1/\sqrt{n})$ ensures that the trade-off term is of the same order as the main term, and we obtain a rate of convergence of order $O(p/n)$ for $\|\hat{\theta} - \theta^*\|$, which is what we can expect from such a model without any sparsity assumption. The second part of this theorem ensures the topological consistency of the reconstructed function $\hat{f}$, namely that it has a persistence that remains close to the persistence of the regression function $f^\star$ and is therefore topologically smooth to some extent. Again choosing $\mu = O(1/\sqrt{n})$ (as suggested above to keep a classical convergence rate for the parameter) leads to a consistency of the persistence of $\hat{f}$ towards that of f^* at a rate $O(p/\sqrt{n})$. We therefore need a larger sample size (of order at least $O(p^2)$) to obtain consistency of the total persistence.

Note that according to Equation 6.14 from [PRSZ20], the number of features with persistence larger than some given value c can be upper bounded by $(\kappa\|\nabla f\|_\infty/c)^d$ where κ is a constant that depends only on the metric of the manifold. This shows a connection between the topological smoothness developed in this section and a more usual notion of smoothness.

III.4.3 Theoretical prospects

A first possible extension of the theoretical results presented in this section can be to generalize Theorem III.6 to mis-specified models, that is, cases where the target function $f^\star$ no longer belongs to $\mathrm{Span}(\Phi_1, \ldots, \Phi_p)$ but can be any L^2 function. To this end, a lower bound on the bias incurred with such a model for a function with fixed total persistence may be found in Proposition 2.1.1 from [PPS19] in the case of surfaces.

Theorem III.7. *Let $\mathcal{M}$ be a compact orientable Riemannian surface without boundary. Denote by $\mathcal{F}_\lambda$ the set of smooth functions over $\mathcal{M}$, such that for every $f \in \mathcal{F}_\lambda$, $\|f\|_2 = 1$ and $\|\Delta f\|_2 \leq \lambda$. Then there exists a constant κ that only depends on $\mathcal{M}$ and its metric such that for every Morse function $f : \mathcal{M} \to \mathbb{R}$,*

$$\inf\{\|f-h\|_\infty | h \in \mathcal{F}_\lambda\} \geq \frac{1}{2(\nu(f)+1)}\left(\mathrm{Pers}(f) - \kappa(\lambda+1)\right).$$

This means that for a fixed λ, if the persistence of the target function f is too large, it will be impossible to approximate it with eigenfunctions of corresponding eigenvalue smaller than λ, and in order to have a chance of approximating it, we will have to allow for more oscillating functions by letting λ increase. Balancing the estimation and approximation errors, based on the above result might lead to a data-driven choice for selecting p.

The other possibility of extension would be to establish consistency results for the case of estimated eigenfunctions, based on the graph Laplacian approach. For example, in order to derive an oracle inequality similar to that of Theorem III.6, we would need to establish the convergence of the persistence of the eigenvectors towards the persistence of the eigenfunctions. A potential approach is to leverage the stability results for persistence (for example, Lemma III.2), and combine them with error rates for eigenfunction estimation (for example, recent results by [DWW21] and [CGTL21] for the empirical uniform norm). However, there are several technical challenges to overcome, in order to implement the above proof strategy. For example, it is not clear if the estimated eigenfunctions satisfy the regularity conditions required, for example, by stability results like Lemma III.2. Furthermore, the results from [DWW21] and [CGTL21] are existence results, and do not resolve the inherent identifiability issue arising in eigenfunction estimation. In addition, the following two cases are to be distinguished:

- Semi-supervised setting: This corresponds to the case when there is an additional set of unlabeled observations $(X_{n+1}, \dots, X_{n+m})$ uniformly and independently sampled over $\mathcal{M}$. In this case, the eigenfunctions could first be estimated using the above unlabeled observations and the regression coefficients could be subsequently estimated using the estimated eigenfunctions.
- Supervised setting: In this case, the same set of observations $(X_1, \dots, X_n)$ is used to estimate the eigenfunctions and the regression coefficients. Extra difficulties arise in this case due to the dependency in estimating the eigenfunctions and the regression coefficients using the same set of observations. We remark that this is the approach we take in the experiments as we discuss in Section III.5.

III.5 Experimental results

III.5.1 Experimental design

We have applied the following experimental routine in order to estimate the function $f^\star$, given a set of points $(X_i)_{i=1}^n$ in $\mathbb{R}^D$ that are assumed to lie on a manifold $\mathcal{M}$ of dimension d and a vector of real responses $(Y_i)_{i=1}^n$.

- Build a proximity graph on the data points X_i. Many options are possible, in most experiments we have taken a $k-$nearest neighbor graph with $k \simeq \log(n)$ but we also sometimes consider Gaussian weighted graphs, for instance in Section III.5.3.
- Compute the normalized Laplacian matrix of the graph.
- For a fixed $p \leq n$, compute the first p eigenvectors of the Laplacian matrix, which yields a new design matrix in $\mathbb{R}^{n \times p}$ where each column is an eigenvector.
- For the penalty Ω_1, compute the persistence of each eigenvector, divide each column of the design matrix by its persistence and solve a Lasso with cross-validation.
- For the penalty Ω_2, we start from a random vector $\theta_0 \in \mathbb{R}^p$ and perform a stochastic gradient descent of $\mathcal{L}$, where we compute the persistence of $\sum_{i=1}^p \theta_i \hat{\Phi}_i$ at each iteration, similarly to what is done in [CCG+21].

The method that has been the most efficient is to take $p = n$ for Ω_1 to perform a variable selection, using Lasso sparsity properties. We then perform a gradient descent of the loss function with penalty Ω_2 on the subset of eigenvectors previously selected. The "vanilla" penalty Ω_2 (without pre-selection step) is itself numerically outperformed in terms of MSE by Ω_1. This can be explained by Theorem III.6 : without performing a preliminary variable

selection, the dimension of the problem does not guarantee a good reconstruction of the source function. This dimension reduction also offers a modest acceleration of the computational cost for the optimization of the loss with penalty Ω_2 as we will see in Section III.5.4. In the tables below, the results for the penalty Ω_2 are always understood to have been obtained this way.

The routine previously described works for the denoising problem where we have a label for each data point. If we are interested in prediction problems, where the set of covariates is split in two: one part with labels and one part without, we build the graph on all the data points since all the points are assumed to lie on $\mathcal{M}$ but we train the model only on the points for which we have a label at our disposal.

Numerically, to compute the persistence of a function f for which we know the values at points $x_1, \dots, x_n$, we build the alpha-complex introduced in [EH22] on the vertex set $(x_1, \dots, x_n)$ where the filtration value of a k-dimensional simplex $\{x_{i_0}, \dots, x_{i_k}\}$ is equal to $\max_{i \in \{i_0, \dots, i_k\}} f(x_i)$. For the alpha-complex to convey the same topology as the underlying manifold, we truncate it by keeping only simplices whose circumradius is small enough (namely smaller than half the reach of the manifold, see e.g., [BCY18]). This reach parameter is known in the simulations of Section III.5.2. Should it not be known, it could be estimated (for instance, using methods developed by [BHHS22]). In the real data examples of Section III.5.3, the underlying simplicial complex is the extension of the nearest-neighbor graph used to compute the Laplacian eigenbasis. We claim that if the number of neighbors is chosen with care, we will also retrieve the topology of the underlying manifold. If using Gaussian-weighted graphs, we introduce a proximity parameter which is equivalent to considering a truncated Rips-filtration. Note that this simplicial complex is only computed once, and does not impact the overall computational cost, even when the ambient dimension is high. Here, we have used the `Gudhi` library [MBGY14] to compute the persistence given this filtration.

In what follows, we will present the results of several experiments conducted both on synthetic and real data to investigate the relevance of our approach in practice. We compare our method to standard regression methods on manifolds: Kernel Ridge Regression (KRR), Nearest-neighbour regression (k-NN), Total variation penalty (TV) as well as graph Laplacian eigenmaps with a L^1 penalty (Lasso) and a weighted Lasso where the weights are the total variation of each eigenvector, computed on the graph (Lasso-TV). The performance is measured in terms of root mean squared error between the estimated and the true functions at the data points. Its expression is given by

$$RMSE(\hat{f}) = \sqrt{\sum_{i=1}^{n} \frac{(f^\star(x_i) - \hat{f}(x_i))^2}{n}}.$$

All hyperparameters have been tuned by cross validation or grid-search. The code used for data on a Swiss roll and the spinning cat in 2D is available here.[2]

III.5.2 Simulated data

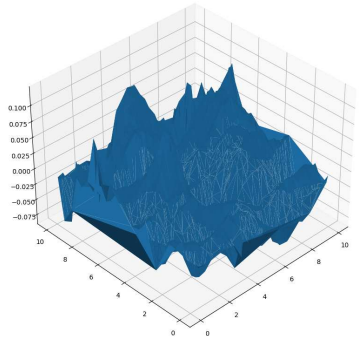
Illustrative example In order to illustrate the behavior of the Ω_2-penalization model, we first look at a synthetic setting where the function we try to reconstruct is a sum of 4 Gaussians, to which we add a large noise (Figure 21).

The persistence diagram of the true function to be estimated has four points in its 1-homology corresponding to the 4 cycles in each Gaussian, all being born at a neighboring

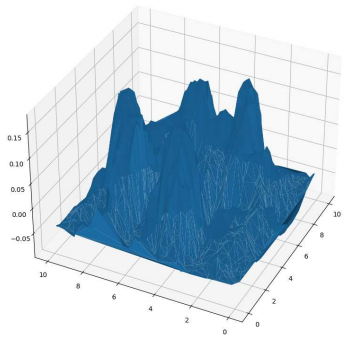
[2] https://github.com/OlympioH/Lap_reg_topo_pen

saddle and dying at the corresponding local maximum, and one point in its 0-homology dying at infinity corresponding to the only connected component. The 0-homology points on the diagonal correspond to sampling noise. When noise is added, the visual chaos of the observation is supported by its persistence diagram: it has a lot of features and the statistical noise added to the measurement here is converted into topological noise.

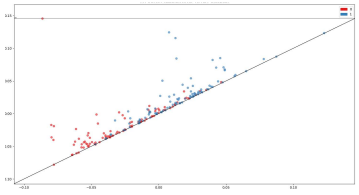
In Figure 25 we compare the results of a stochastic gradient descent penalizing Ω_2 against a Lasso estimation. The reconstruction is visually better and is also better in terms of MSE. Indeed, when penalizing the persistence, we have managed to keep the four most persistent one-dimensional features in the persistent diagram (although their persistence has diminished) and we still observe four peaks, while most of the noise has been removed. Although the Lasso enables some denoising, it has only been able to reconstruct 2 to 3 peaks and does not offer the same denoising performance. Note that it is also possible to consider a topological penalty where we penalize the persistence of the function except the 4 most persistent points in 1-homology and the most persistent point in 0-homology like in [BGNDS20]. In this case, we can be more coarse in the choice of the trade-off μ since the penalty Ω_2 must then be set to 0. This shows that if one is given an a priori on the topology of the function to reconstruct, it can be included in the model very easily.



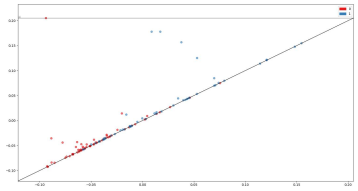
(a) Function estimated by a Lasso



(b) Function estimated by the topological penalty Ω_2



(c) Persistence diagram of the Lasso estimate



(d) Persistence diagram of the Ω_2 penalty estimate

Figure 25: Reconstruction of the sum of four Gaussians.

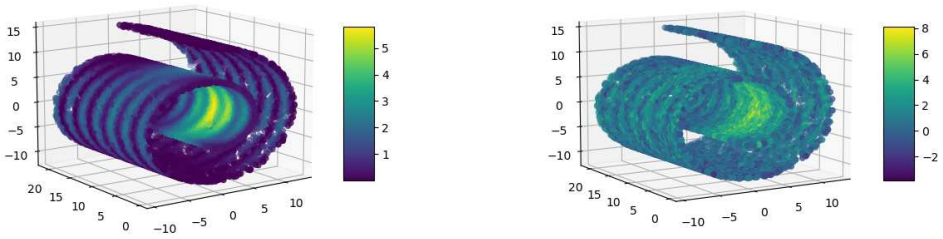
Torus data We simulate data on a torus which we recall is homeomorphic to $\mathbb{S}^1 \times \mathbb{S}^1$, parametrized by the embedding $\Psi_{\mathbb{T}^2} : (\theta, \varphi) \mapsto (x_1, x_2, x_3)$:

$$\begin{cases} x_1 = (2 + \cos(\theta)) \cos(\varphi), \\ x_2 = (2 + \cos(\theta)) \sin(\varphi), \\ x_3 = \sin(\theta). \end{cases}$$

Following the approach of [DHS+13], sampling uniformly on the torus may be carried out

Table 2: RMSE of the reconstruction for n points lying on a torus, average on 100 runs.

n	σ	Lasso	Lasso-TV	Ω_1	Ω_2	KRR	k-NN	TV
300	0.5	0.261 ± 0.019	0.241 ± 0.017	0.220 ± 0.019	0.223 ± 0.019	**0.217 ± 0.012**	0.237 ± 0.017	0.413 ± 0.016
300	1	0.381 ± 0.038	0.337 ± 0.043	**0.281 ± 0.038**	0.288 ± 0.037	0.292 ± 0.025	0.401 ± 0.029	0.509 ± 0.026
1000	0.5	0.174 ± 0.011	0.171 ± 0.010	0.157 ± 0.010	**0.156 ± 0.011**	0.172 ± 0.008	0.167 ± 0.009	0.421 ± 0.008
1000	1	0.290 ± 0.024	0.266 ± 0.021	0.212 ± 0.018	**0.209 ± 0.018**	0.222 ± 0.017	0.285 ± 0.013	0.513 ± 0.015



(a) Original function (b) Noisy observation

Figure 26: Regression function on a Swiss roll.

via sampling φ uniformly in $[0, 2\pi]$ and θ according to the density $g(\theta) = \frac{1}{2\pi}\left(1 + \frac{\cos(\theta)}{2}\right)$ on $[0, 2\pi]$. In practice, this is performed by a rejection sampling. Note that sampling θ and φ uniformly and independently does not provide a uniform sampling on the torus as we will observe a higher density of points in the inside of the torus. A function on the torus is identified with a function of (θ, φ). To set up the regression problem, we define the target response $f^*(\theta, \varphi) = \xi[-17(\sqrt{(\theta - \pi)^2 + (\varphi - \pi)^2} - 0.6\pi)]$ where ξ is the sigmoid function. Note that f^* is radial symmetric, depending only on the distance between (θ, φ) and (π, π). This signal function has been studied because it has a simple topology that illustrates the topological denoising method. It has first been introduced by [NSJ07] for similar purposes. The comparison of the RMSE for all methods can be found in Table 2.

Data on a Swiss roll We now consider data on a Swiss roll which is a two dimensional manifold parametrized by the mapping $(x, y) \mapsto (x \cos x, y, x \sin x)$. We have set as target function

$$f^*(x, y) = 4 \exp(-((y - 7)^2/20 + (x - 6)^2/5)) + 2 \cos^2(x) \sin^2(y),$$

for $(x, y) \in [1.5\pi, 3.5\pi] \times [0, 21]$. The function is plotted Figure 26.

Here, we observe that the penalty Ω_2 performs the best and benefits from the selection properties of the regularization with penalty Ω_1. Data on a Swiss roll shows the limitations of Kernel methods when the number of points is low, as the geodesic metric can be very different from the ambient metric. The use of a $k-$NN graph is a solution to bypass this problem since at a small scale the two metrics are close. We see in Table 3 that a topology-based penalty is particularly efficient when the noise level becomes important. Although Kernel Ridge regression and Total Variation penalties provide a very good reconstruction and should be preferred when the noise is low, they become quite unstable as the noise level

Table 3: RMSE of the reconstruction for 500 points on a Swiss roll, average on 100 runs.

σ	Lasso	Lasso - TV	Ω_1	Ω_2	KRR	k-NN	TV
0.2	0.422 ± 0.035	0.421 ± 0.031	0.491 ± 0.065	0.398 ± 0.026	**0.186 ± 0.006**	0.505 ± 0.020	0.200 ± 0.006
0.5	0.498 ± 0.036	0.478 ± 0.026	0.472 ± 0.043	**0.455 ± 0.020**	0.476 ± 0.014	0.532 ± 0.020	0.494 ± 0.017
0.7	0.549 ± 0.041	0.525 ± 0.033	0.534 ± 0.032	**0.489 ± 0.021**	0.526 ± 0.017	0.549 ± 0.021	0.671 ± 0.022
1	0.628 ± 0.0430	0.593 ± 0.035	0.572 ± 0.0600	**0.546 ± 0.030**	0.637 ± 0.026	0.587 ± 0.023	0.920 ± 0.029
1.3	0.689 ± 0.049	0.648 ± 0.042	0.615 ± 0.064	**0.595 ± 0.038**	0.746 ± 0.026	0.646 ± 0.025	1.056 ± 0.043

(a) Baseline image (b) Image rotated by 30° (c) Image rotated by 90°

Figure 27: Data set for the experiments on a one dimensional manifold with real data.

increases, whereas all Laplacian eigenmaps based methods as well as a k-NN regression are somehow robust to noise. Note that among all possible penalties on Laplacian eigenmaps based models, Ω_2 ran on the eigenfunctions selected by Ω_1 always yield the best results.

III.5.3 Real data

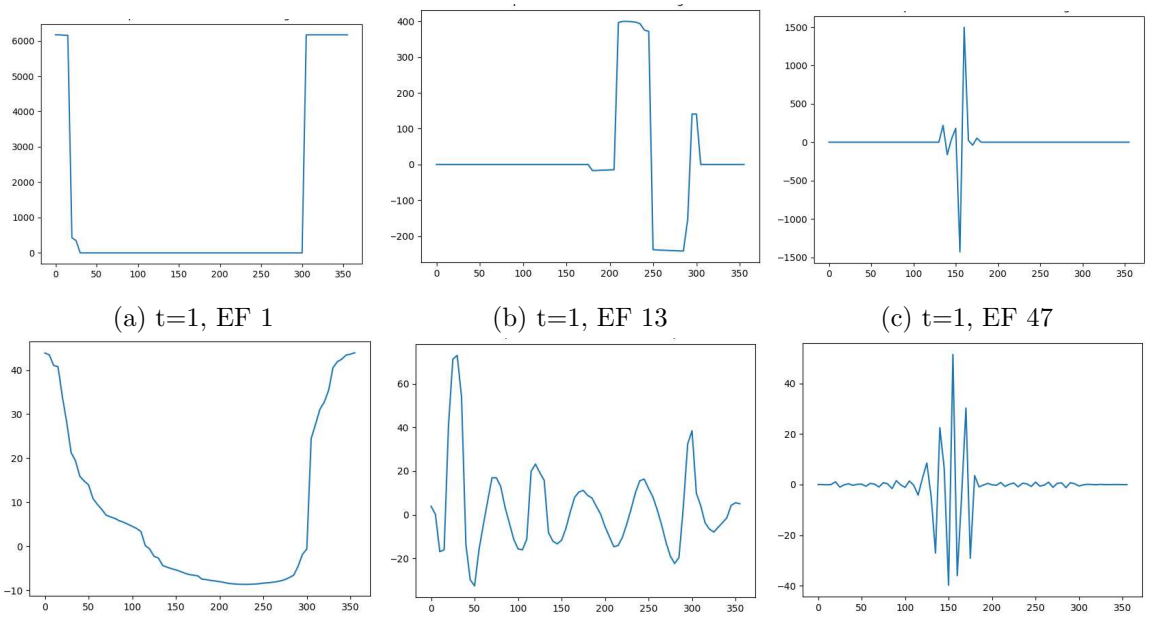
Spinning cat in 1D This model has also been tried on real data. We consider a data set of $n = 72$ images of the same object rotated in space with increments of 5°. The data lie on a one-dimensional submanifold of $\mathbb{R}^{16384}$, the images having size 128×128 pixels. We can see some of the images from the dataset Figure 27. For each image, we want to retrieve the angle of rotation in radians of the object. The source vector we want to estimate is therefore $(0, 5\pi/180, 10\pi/180, \ldots, 355\pi/180) \in \mathbb{R}^{72}$.

We have built a Gaussian weighted graph on the data points, using the ambient L^2 metric between images for the weights. Using a geodesic distance between images would probably yield better results, but the use of the ambient metric is enough to have convergence of the graph Laplacian eigenvectors towards the eigenfunctions of the Laplace-Beltrami operator on the manifold according to [GTGHS20].

We have tried different values of the scaling parameter t in the Gaussian weights

$$W_{ij} = \frac{1}{n} \frac{1}{t(4\pi t)^{d/2}} e^{-\frac{\|x_i - x_j\|^2}{4t}}.$$

We depict in Figure 28 the aspect of some eigenfunctions as a function of the rotation degree of the baseline image for $t = 1$ and $t = 10$. Here, we can see that the value $t = 1$ is too small: indeed, although the graph-Laplacian eigenvectors converge to the true Laplace-Beltrami eigenfunctions as $n \to \infty$ and $t \to 0$, this only occurs if t verifies a particular scaling



(a) t=1, EF 1 (b) t=1, EF 13 (c) t=1, EF 47

(d) t=10, EF 1 (e) t=10, EF 13 (f) t=10, EF 47

Figure 28: Estimated eigenfunctions of the manifold Laplacian evaluated at each image, as functions of the degree of rotation of the object in the image.

Table 4: RMSE of the prediction of the angle of rotation of the spinning cat in 1D, average on 100 runs.

σ	Size train/test	Lasso	Lasso-TV	Ω_1	Ω_2	KRR	k-NN
0	68/4	0.79 ± 0.59	0.56 ± 0.52	0.38 ± 0.51	0.35 ± 0.49	**0.22 ± 0.47**	0.25 ± 0.50
0.5	68/4	0.89 ± 0.60	0.70 ± 0.51	0.59 ± 0.52	0.53 ± 0.48	**0.47 ± 0.47**	0.44 ± 0.47
0	60/12	0.96 ± 0.50	0.74 ± 0.46	0.58 ± 0.47	0.54 ± 0.46	**0.45 ± 0.41**	0.65 ± 0.41
0.5	60/12	1.04 ± 0.56	0.85 ± 0.50	0.71 ± 0.48	0.66 ± 0.49	**0.57 ± 0.42**	0.77 ± 0.44

with respect to n according to [GTGHS20]. Here, we only have a small number of data at hand ($n = 72$), and therefore, the value $t = 10$ visually seems to be more satisfying. Indeed, the data are on a circle (of some large dimensional Euclidean space) and we would expect the eigenfunctions to converge towards the spherical harmonics for the circle, which are oscillating functions. For $t = 1$, all the eigenfunctions are highly localized which is not quite satisfying. In addition, a preliminary study yielded a much better performance of the parameter $t = 10$ over $t = 1$ for the corresponding regression task. For a larger index, the eigenfunctions start to be localized around an image of the manifold, reminding a wavelet-type basis.

Unlike the synthetic experiments, where we have only focused on denoising, here we are interested in the prediction properties of the method. To this aim, the data are randomly split between a training set and a validation set. The graph is then built on the whole dataset (since the data points are all assumed to lie on the same manifold). The optimization procedure is then only performed on the labels from the training set, and we measure the mean square error between the prediction on the new points and the true values from the validation set.

We see here that although an approach based on the penalization of the topological persis-

Table 5: RMSE of the prediction, regression on a 2D manifold with real data, average of 100 runs.

train/test	Lasso	Lasso-TV	Ω_1	Ω_2	KRR	k-NN
900/100	0.346 ± 0.031	0.344 ± 0.036	0.302 ± 0.032	0.291 ± 0.029	**0.281 ± 0.027**	0.312 ± 0.035
800/200	0.351 ± 0.033	0.317 ± 0.030	0.283 ± 0.028	**0.273 ± 0.029**	0.290 ± 0.025	0.309 ± 0.024
700/300	0.348 ± 0.028	0.314 ± 0.029	0.293 ± 0.025	**0.280 ± 0.024**	0.296 ± 0.023	0.320 ± 0.023
600/400	0.344 ± 0.027	0.306 ± 0.025	0.304 ± 0.028	0.297 ± 0.027	**0.290 ± 0.017**	0.333 ± 0.024

tence yields results a lot better than a L^1 or total variation penalty, it is still outperformed by a kernel ridge regression. This might be due to the small number of data available on which to build the graph or the fact that the topology of the manifold as well as the topology of the source function are very simple and do not benefit fully from the method presented in this section. We will see in the next subsection a set-up that shows the appeal of such a penalty.

Spinning cat in 2D We consider the data set from the previous subsection where in addition each image is rotated in another direction by increments of 5°. We thus dispose of a two-dimensional manifold with the homotopy type of a torus, where the two parameters are the degree of rotation of the object within the image θ and the degree of rotation of the image itself φ. A few images of the dataset are plotted in Figure 29. We consider the same target response as in the Subsection III.5.2:

$$f^*(\theta, \varphi) = \xi[-17(\sqrt{(\theta-\pi)^2 + (\varphi-\pi)^2} - 0.6\pi)].$$

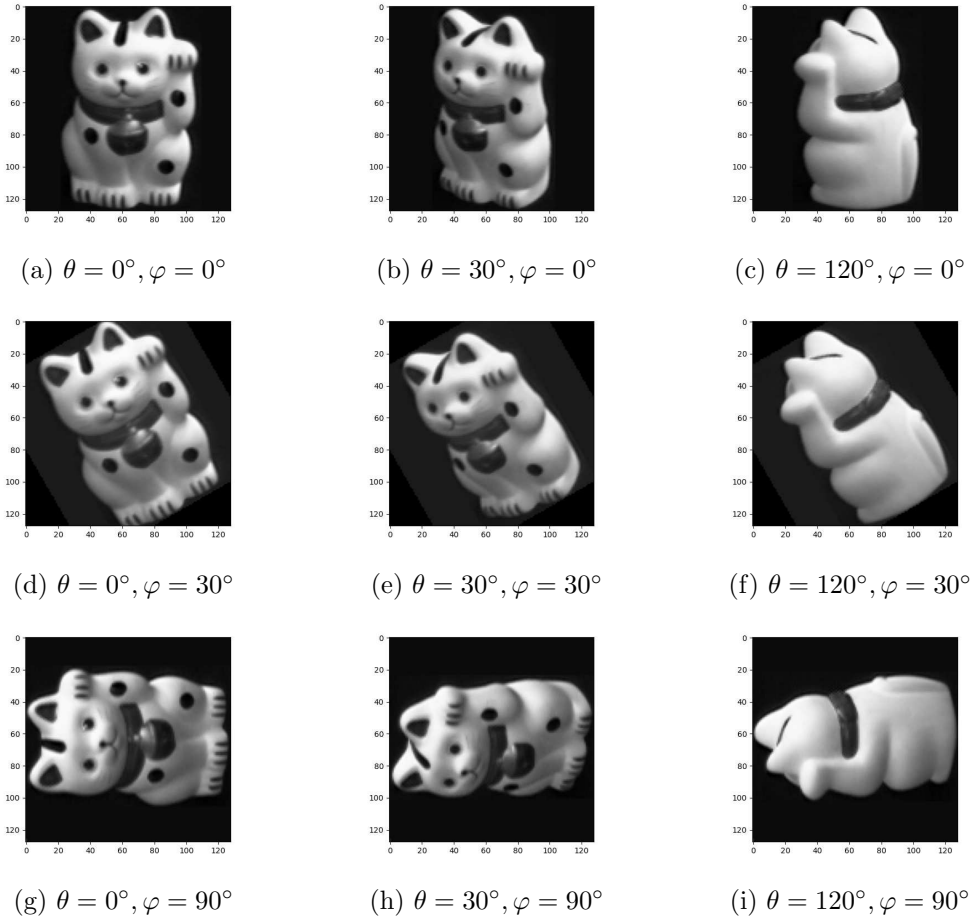
We add an i.i.d. Gaussian noise with standard deviation $\sigma = 1$ to the input. We select a random subset of 1000 images over the dataset and randomly split it into a training set and a testing set. In this example, we only penalize the $0-$persistence since it has provided a better performance on a preliminary study. The results can be found Table 5.

Among all prediction methods that make use of a Laplacian eigenbasis decomposition, Ω_2 on a subset of eigenvectors preselected thanks to Ω_1 offers the best performance when predicting to new data. Our method is overall comparable to Kernel Ridge Regression.

Electrical consumption dataset We have tried our method on the electrical consumption dataset.[3] The covariates are curves of temperature in Spain, averaged over a week. There is a measurement for each hour of the day, thus each curve can be seen as a point of $\mathbb{R}^{24}$. However, the possible profiles of temperature are very limited (namely, each curve is increasing towards a maximum in the afternoon and is then decreasing), it is therefore believed that the data lie on a manifold of smaller dimension. For each week, we try to predict the electrical consumption in Spain in GW. Like in the previous experiments, the data are randomly split between a training and a testing set, and the graph Laplacian is built on all the available data. The results can be found Table 6.

On this dataset, we notice once again that a small training set is not really damaging towards topological methods as opposed to more standard methods, illustrating once again the generalization properties of the penalties Ω_1 and Ω_2. Here, all methods act relatively similarly as they all predict well the electrical consumption over a week when it takes values around its mean (see Figure 30). However, sometimes the electrical consumption over a week

[3] https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather



(a) $\theta = 0°, \varphi = 0°$ (b) $\theta = 30°, \varphi = 0°$ (c) $\theta = 120°, \varphi = 0°$

(d) $\theta = 0°, \varphi = 30°$ (e) $\theta = 30°, \varphi = 30°$ (f) $\theta = 120°, \varphi = 30°$

(g) $\theta = 0°, \varphi = 90°$ (h) $\theta = 30°, \varphi = 90°$ (i) $\theta = 120°, \varphi = 90°$

Figure 29: Dataset of images on a 2-dimensional manifold.

reaches a peak and all methods fail to capture the extreme values of the source function. It makes perfect sense here that the knowledge of the temperature alone fails to explain perfectly the global electrical consumption over an entire country, without taking into account any other covariates.

III.5.4 Discussion on the computational cost

The good performance of topological penalties, both in terms of prediction and reconstruction have to be nuanced by their computational cost. Table 7 shows the computational time in seconds on a standard laptop without GPU for the example of Section III.5.2 (radial peak function on a torus). All the methods have been implemented in Python using standard libraries, except for Ω_2 and TV for which the optimization of the loss function has been

Table 6: RMSE of the prediction of the average electrical consumption.

train/test	Lasso	Lasso-TV	Ω_1	Ω_2	KRR	$k-$NN
108/100	1.289 ± 0.067	1.216 ± 0.065	1.174 ± 0.072	1.251 ± 0.063	**1.168 ± 0.072**	1.188 ± 0.064
58/150	1.254 ± 0.044	1.265 ± 0.050	1.181 ± 0.048	**1.165 ± 0.046**	1.193 ± 0.051	1.211 ± 0.059

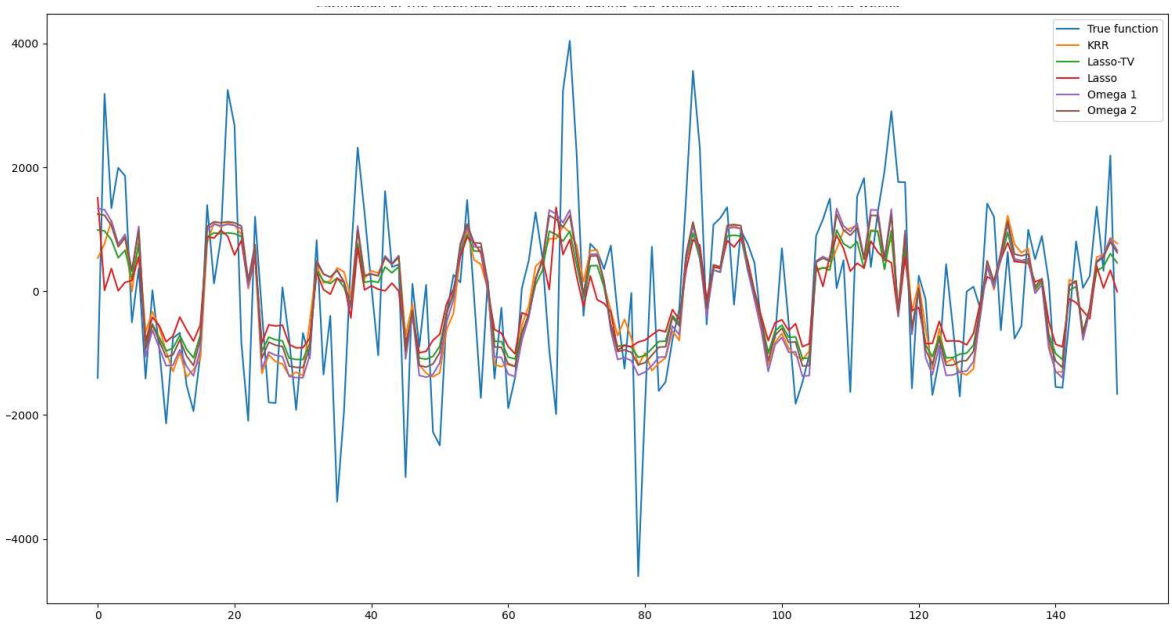


Figure 30: Prediction of the electrical consumption using various methods.

implemented from scratch. This explains a higher cost as opposed to standard regression methods that already benefit from an optimized implementation. This table presents the cost for various methods for single-choices of parameters and without any cross validation. We observe that computing the graph Laplacian and its spectrum is very fast (less than a second if we have 1000 points and ask for the entire spectrum), and performing a Lasso has a negligible cost. Most of the time is actually spent computing persistence diagrams : for the penalty Ω_1, we have to compute the persistence of each eigenfunction prior to performing a standard Lasso. This turns out to be very costly if we ask for the whole spectrum and somehow reasonable if we ask for a small value of p. Note that once these persistences have been computed, estimating a new signal at the same data points can be done almost instantly. Penalty Ω_2 has a very high computational cost, which does not decrease significantly with the dimension p. This is due to the fact that the persistence diagram of the entire function has to be computed at each gradient step.

Computing the $k-$persistence of a simplicial complex with N simplices is done with the `Gudhi` library [MBGY14] which relies on the algorithm of [EH22] and has an algorithmic complexity of $O(N^3)$. Note that if we have n data-points, the number of $k-$dimensional simplices is of order n^k. This accounts for the very high computational cost of topological penalties developed in this section. It is therefore infeasible to use this method in practice for high homological dimensions, namely as soon as $k \geq 4$; in practice we recommend to only penalize k-persistences, for k up to 3. On the other hand, note that the computation time is barely impacted by the intrinsic dimension of the manifold nor the ambient dimension. We remark that when minimizing Ω_2, the function does not change much from one epoch to another and therefore, neither does its persistence diagram. Recomputing the diagram at each iteration is therefore a naive approach and this is a possible way of improving the method on the numerical side. We believe that using vineyards [CSEM06] would enable a computation of the persistence diagram in linear time which would speed up the method, maybe at the

Table 7: Computational time (in seconds), data on a torus

n	p	Lasso	Ω_1	Ω_2	KRR	k-NN	TV
300	100	0.04	4.0	36.3	0.11	3.0×10^{-3}	2.3
1000	100	0.19	14.2	129.9	0.13	5.6×10^{-3}	3.6
300	300	0.11	8.5	39.7	0.11	3.0×10^{-3}	2.3
1000	1000	0.94	132.3	195.9	0.13	5.6×10^{-3}	3.6

cost of a higher space complexity.

III.5.5 Conclusion of the experiments

The methods developed in this section couple the use of a graph Laplacian eigenbasis and a topological penalty. The first aspect enables to treat regression problems for data living on manifolds with an extrinsic approach: nothing needs to be known on the manifold and its metric, the graph Laplacian being computed on the ambient metric. Laplace eigenmaps methods in general have proven to be useful when the manifold structure is quite strong, illustrated here with the experiment on a Swiss roll. The use of a topological penalty has multiple advantages: it acts as a generalization of total variation penalties in higher dimensions and seems to be a more natural way to regularize functions, as observed in Section III.2. Numerically, topological methods almost always provide better results than usual penalties such as TV or L^1 penalty. Here, we have developed two types of topological penalties: Ω_1 aims at performing a selection process of the regression basis functions, by discarding the ones that oscillate too much in order to allow a good generalization to new data. On the other hand Ω_2 directly acts on the topology of the source function and performs a strong denoising. Note that the statistical noise on the data translates into a topological noise on the persistence diagram of the corresponding function which is the one the regularization Ω_2 acts onto, by providing a powerful simplification of the persistence diagram of the reconstructed function. When the underlying geometric structure is complex or the noise is important, Ω_2 regularization on the eigenfunctions selected by Ω_1 provides a better reconstruction in terms of RMSE than standard methods, including KRR which appears to be the most competitive one.

We have essentially penalized the total persistence of functions, but it would be possible to numerically penalize any smooth function on the points of the persistence diagram. In particular, establishing a penalty that would erase all low-dimensional persistence features while keeping all the high-dimensional features could be of practical interest, likewise to the smoothly clipped absolute deviation (SCAD) penalty developed by [FL01].

One major drawback of topological methods is their computational cost as opposed to a simple KRR regression, especially when we need to compute topological persistence of high dimensions.

III.6 Proofs for Section III.4

This section is devoted to the proofs of the main theoretical results of this section.

III.6.1 Proof of Theorem III.6

We start by the proof of Theorem III.6 since it introduces a number of lemmas and results that will also be useful to prove the other theorems. We start with a technical lemma based on the celebrated Weyl's Law [Cha84, Ivr16], reformulated for statistical regression purposes in [BBM99]. It provides a control of the sup-norm of the sum of the squares of the first Laplacian eigenfunctions.

Lemma III.8. *Let $\mathcal{M}$ be a compact Riemannian manifold of dimension d and Δ its Laplace-Beltrami operator. Let $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p \leq \ldots$ be the eigenvalues of Δ, and for an eigenvalue λ_i, we denote by Φ_i a normalized eigenfunction. Then there exists constants $C'(\mathcal{M})$ and $C(\mathcal{M})$ depending only on the geometry of $\mathcal{M}$ (with $C(\mathcal{M})$ also potentially depending on the dimension) such that for all p,*

$$\left\| \sum_{i=1}^{p} \Phi_i^2 \right\|_\infty \leq C'(\mathcal{M}) \lambda_p^{d/2} \leq C(\mathcal{M}) p,$$

where the last inequality follows by Weyl's law.

We will also need the following concentration result, itself using Lemma III.8. Recall that a random variable ϵ is called sub-Gaussian with parameter $\sigma^2 > 0$ if $\mathbb{E}\exp(\lambda\epsilon) \leq \exp(\sigma^2\lambda^2/2)$ for all $\lambda \in \mathbb{R}$.

Lemma III.9. *Denote for every $X \in \mathcal{M}$, the tuple $\Phi(X) = (\Phi_1(X), \ldots, \Phi_p(X))$. Let $\varepsilon_1 \ldots, \varepsilon_n$ be i.i.d sub-Gaussian random variables with parameter σ^2, independent of $X_1, \ldots, X_n$, and let $x > 0$. Then, with probability larger than $1 - 2e^{-x}$,*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \Phi(X_i) \right\| \leq 2\sigma \sqrt{\frac{p}{n}} (1 + 2\sqrt{x}) \sqrt{1 + C(\mathcal{M}) \sqrt{\frac{2x}{n}}}.$$

Proof. (of Lemma III.9). Our proof is based on a non-standard result by [Kon14], on concentration of Lipschitz functions of not necessarily bounded random vectors, which we briefly introduce next.

For $i = 1, \ldots, n$, let Z_i be random objects living in a measurable space $\mathcal{M}_i$ with metric ρ_i, endowed with measure μ_i. Define $Z = (Z_1, \ldots, Z_n)$ to be living on the product probability space $\mathcal{M}_1 \times \cdots \times \mathcal{M}_n$ with product measure $\mu_1 \times \cdots \times \mu_n$, and metric $\rho = \sum_{i=1}^n \rho_i$. To each (Z_i, μ_i, ρ_i), we also associate symmetrized random objects $\Xi_i = \gamma_i \rho_i(Z_i, Z_i')$, where $Z_i, Z_i' \sim \mu_i$ are independent, and γ_i are Rademacher random variables (i.e. taking values ± 1 with probability $1/2$) independent of Z_i, Z_i'. We also define $\Delta_{\mathrm{SG}}(\Xi_i)$ as the sub-Gaussian diameter of Ξ_i, given by the smallest value of s for which we have $\mathbb{E}e^{\lambda \Xi_i} \leq e^{s^2\lambda^2/2}$ for all $\lambda \in \mathbb{R}$. Then using [Kon14, Proof of Theorem 1], we have for the random object Z as above, and a 1-Lipschitz function $\varphi : \mathcal{M}_1 \times \cdots \times \mathcal{M}_n \to \mathbb{R}$, that

$$\mathbb{P}(\varphi(Z) - \mathbb{E}\varphi(Z) > t) \leq \exp\left(- \frac{t^2}{2 \sum_{i=1}^n \Delta_{\mathrm{SG}}^2(Z_i)} \right). \tag{III.7}$$

In our context, for all $1 \leq i \leq n$, we set $\mathcal{M}_i = \mathbb{R}^p$, and $\rho_i = \|\cdot\|$. For $v_i \in \mathbb{R}^p$, we define the function φ as

$$\varphi : (v_1, \ldots, v_n) \mapsto \left\| \sum_{i=1}^{n} v_i \right\|.$$

Note that, for two finite sets of vectors $(v_i)_i$ and $(w_i)_i$, we have

$$\left| \left\| \sum v_i \right\| - \left\| \sum w_i \right\| \right| \leq \sum \|v_i - w_i\|.$$

Hence, the function φ is $1-$Lipschitz with respect to the mixed ℓ^1/ℓ^2 norm of $(\mathbb{R}^p)^n$, meaning that we will be able to apply Kontorovich's result.

Now, considering $\left\|\sum_{i=1}^n \varepsilon_i \Phi(X_i)\right\|$, we proceed by conditioning on $(X_i)_{1\leq i\leq n}$, that is, we consider randomness only with respect to $(\varepsilon_i)_{1\leq i\leq n}$, which are independent of the X_i's; $\mathbb{E}_\varepsilon$ will denote the corresponding conditional expectation. First, note that we have the following bound on the (conditional) expectation:

$$\mathbb{E}_\varepsilon\left(\left\|\sum_{i=1}^n \varepsilon_i \Phi(X_i)\right\|\right) \leq \mathbb{E}_\varepsilon\left(\left\|\sum_{i=1}^n \varepsilon_i \Phi(X_i)\right\|^2\right)^{\frac{1}{2}} \leq \sqrt{\sigma^2 \sum_{i=1}^n \|\Phi(X_i)\|^2}. \tag{III.8}$$

Defining $Z_i = \varepsilon_i \Phi(X_i)$, and considering the corresponding symmetrized object

$$\Xi_i = \gamma_i \left\|\varepsilon_i \Phi(X_i) - \varepsilon_i' \Phi(X_i)\right\| = \gamma_i |\varepsilon_i - \varepsilon_i'| \|\Phi(X_i)\|,$$

and since $\gamma_i|\varepsilon_i - \varepsilon_i'|$ has the same distribution as $(\varepsilon_i - \varepsilon_i')$ by independence and symmetry, we have

$$\mathbb{E}_\varepsilon\left(\exp(\lambda \Xi_i)\right) = \mathbb{E}_\varepsilon\left(\exp(\lambda(\varepsilon_i - \varepsilon_i')\|\Phi(X_i)\|)\right) \leq \exp(\lambda^2 \sigma^2 \|\Phi(X_i)\|^2)$$

since the $\varepsilon_i, \varepsilon_i'$ are independent sub-Gaussian of parameter σ^2, hence we also have for the (conditional) sub-Gaussian diameter $\Delta_{\mathrm{SG}}^2(\Xi_i) \leq 2\sigma^2 \|\Phi(X_i)\|^2$. Hence, by using (III.7) and (III.8), we have that with probability larger than $1 - e^{-x}$,

$$\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \Phi(X_i)\right\| \leq 2\frac{\sqrt{\sigma^2 \sum_{i=1}^n \|\Phi(X_i)\|^2 x}}{n} + \frac{\sqrt{\sigma^2 \sum_{i=1}^n \|\Phi(X_i)\|^2}}{n}.$$

We now deal with removing the conditioning on X_i. Note that according to Lemma III.8, we have for all i: $\|\Phi(X_i)\|^2 = \sum_{j=1}^p \Phi_j(X_i)^2 \leq C(\mathcal{M})p$. We can therefore apply Hoeffding's inequality and obtain that with probability larger than $1 - e^{-x}$,

$$\sum_{i=1}^n \|\Phi(X_i)\|^2 \leq np + C(\mathcal{M})p\sqrt{2nx} = np\left(1 + C(\mathcal{M})\sqrt{\frac{2x}{n}}\right),$$

where we used that $\mathbb{E}\|\Phi(X_i)\|^2 = np$, because the eigenfunctions are normalized. We therefore obtain the desired result. □

For a given function f on a probability space $(\mathcal{X}, \pi)$, we define the expectation operator $P(f) = \int f \mathrm{d}\pi$. Given n i.i.d. random variables on $\mathcal{X}$, we define the empirical measure $P_n(f) = \frac{1}{n}\sum_{i=1}^n f(X_i)$. The following lemma provides a control of the empirical process of the difference between the true function and its estimation:

Lemma III.10. *With the same notation as before, we have that with probability larger than* $1 - \exp\left(\frac{-0.1\,n}{C(\mathcal{M})p} + \ln(2p)\right)$*:*

$$\sup_{\|\beta\|=1} (P_n - P)(\langle \beta, \Phi(X)\rangle^2) \leq \frac{1}{2}.$$

Proof. (of Lemma III.10). First note that for the empirical process we have

$$\sup_{\|\beta\|=1} (P_n - P)(\langle \beta, \Phi(X)\rangle^2) = \sup_{\|\beta\|=1} \beta^t \left(\frac{1}{n}\sum_{i=1}^n \Phi(X_i)\Phi(X_i)^T - \mathbb{E}[\Phi(X)\Phi(X)^T] \right) \beta$$
$$= \left\| \frac{1}{n}\sum_{i=1}^n \Phi(X_i)\Phi(X_i)^T - I_p \right\|_{\text{op}},$$

where we have used $\mathbb{E}[\Phi(X)\Phi(X)^T] = I_p$, since the components $(\Phi_i)_{1\leq i\leq p}$ of Φ form an orthonormal system of $L^2(\mathcal{M})$. Note that $\|\Phi(X_i)\Phi(X_i)^T\|_{\text{op}} = \|\Phi(X_i)\|^2$ is upper-bounded by $L = C(\mathcal{M})p$ according to Lemma III.8. We then use Theorem 5.1.1 from [Tro15] which yields that:

$$\mathbb{P}\left(\left\| \frac{1}{n}\sum_{i=1}^n \Phi(X_i)\Phi(X_i)^T - I_p \right\|_{\text{op}} \geq \frac{1}{2} \right) \leq 2pK^{\frac{n}{C(\mathcal{M})p}}. \tag{III.9}$$

A simple calculation and evaluation of the numerical constant $K = \max\left(\frac{e^{-1/2}}{(1/2)^{1/2}}, \frac{e^{1/2}}{(3/2)^{3/2}} \right)$ gives the required result. □

We are now in a position to prove Theorem III.6. For $f_\theta = \sum_{i=1}^p \theta_i \Phi_i$, let the quadratic loss be denoted by

$$\gamma(\theta, (x, y)) = (f_\theta(x) - y)^2 = (\langle \theta, \Phi(x)\rangle - y)^2.$$

We first remark that since the Φ_i's are an orthonormal system, $\mathbb{E}(f_\theta(X) - f_{\theta^\star}(X))^2 = \|\theta - \theta^\star\|^2$. Moreover, it is a well-known property of the quadratic loss that the excess risk of any prediction function f is the squared L^2 distance to the optimal regression function $f^* = f_{\theta^*}$, so that $P(\gamma(\theta, (X, Y))) - P(\gamma(\theta^\star, (X, Y))) = \mathbb{E}(f_\theta(X) - f_{\theta^*}(X))^2 = \|\theta - \theta^\star\|^2$. Since the test data point $(X, Y) \sim P$ used to compute the risk is independent of the sample used to construct the estimator $\hat{\theta}$, we also have $P(\gamma(\hat{\theta}, (X, Y))) - P(\gamma(\theta^\star, (X, Y))) = \|\hat{\theta} - \theta^*\|^2$ (conditionally on the training sample).

Since $\hat{\theta}$ is a minimizer of $\mathcal{L}$ given in (III.2) with $\Omega_2(\theta) = \text{Pers}(f_\theta)$, we have that

$$P_n(\gamma(\hat{\theta}, .)) - P_n(\gamma(\theta^\star, .)) + \mu\text{Pers}(f_{\hat{\theta}}) - \mu\text{Pers}(f_{\theta^\star}) \leq 0. \tag{III.10}$$

Therefore, we have

$$\begin{aligned}
\|\hat{\theta} - \theta^*\|^2 &= P\gamma(\hat{\theta}, \cdot) - P\gamma(\theta^\star, \cdot) \\
&\leq (P - P_n)(\gamma(\hat{\theta}, \cdot) - \gamma(\theta^\star, \cdot)) + \mu(\text{Pers}(f_{\theta^\star}) - \text{Pers}(f_{\hat{\theta}})). \\
&\leq (P - P_n)(-2(\langle \theta^\star, \Phi\rangle + \varepsilon)\langle \hat{\theta}, \Phi\rangle + \langle \hat{\theta}, \Phi\rangle^2 + 2(\langle \theta^\star, \Phi\rangle + \varepsilon)\langle \theta^\star, \Phi\rangle - \langle \theta^\star, \Phi\rangle^2) \\
&\qquad + \mu(\text{Pers}(f_{\theta^\star}) - \text{Pers}(f_{\hat{\theta}})) \\
&\leq \underbrace{(P_n - P)(2\varepsilon\langle \hat{\theta} - \theta^\star, \Phi\rangle)}_{A} + \underbrace{(P_n - P)(\langle \hat{\theta} - \theta^\star, \Phi\rangle^2)}_{B} + \underbrace{\mu(\text{Pers}(f^\star) - \text{Pers}(\hat{f}))}_{C}.
\end{aligned}$$

We will now bound terms A, B and C below. First note that

$$A \leq 2\|\theta^\star - \hat{\theta}\| \sup_\theta \left\langle \frac{\theta - \theta^\star}{\|\theta - \theta^\star\|}, \frac{1}{n}\sum_{i=1}^n \varepsilon_i \Phi(X_i) \right\rangle \leq 2\|\hat{\theta} - \theta^\star\| \left\| \frac{1}{n}\sum_{i=1}^n \varepsilon_i \Phi(X_i) \right\|,$$

and that Lemma III.9 provides control of the norm on the right-hand side. Next note that according to Lemma III.10, we have

$$B \leq \frac{1}{2}\|\hat{\theta} - \theta^*\|.$$

Finally, to control term C, we simply apply Lemma III.2, which yields

$$\begin{aligned} C = \mu(\mathrm{Pers}(f^\star) - \mathrm{Pers}(\hat{f})) &\leq \mu(2\nu(f^\star) + \zeta)\|\hat{f} - f^\star\|_\infty \\ &= \mu(2\nu(f^\star) + \zeta)\|\langle \hat{\theta} - \theta^*, \Phi(\cdot)\rangle\|_\infty \\ &\leq \mu(2\nu(f^\star) + \zeta)\|\hat{\theta} - \theta^\star\| \left\| \sum_{i=1}^p \Phi_i^2 \right\|_\infty^{1/2} \\ &\leq \mu\sqrt{C(\mathcal{M})}(2\nu(f^\star) + \zeta)\|\hat{\theta} - \theta^\star\|\sqrt{p}, \end{aligned}$$

where the second to last inequality uses the Cauchy-Schwarz inequality, and the last is using Lemma III.8. Finally, we have, with probability larger than $1 - 2e^{-x} - \exp\left(\frac{-0.1n}{C(\mathcal{M})p} + \ln(2p)\right)$,

$$\begin{aligned} \|\hat{\theta} - \theta^\star\|^2 \leq 2\|\hat{\theta} - \theta^\star\|\, 2\sigma\sqrt{\frac{p}{n}}(1+\sqrt{x})\sqrt{1 + C(\mathcal{M})\sqrt{\frac{2x}{n}}} \\ + \frac{1}{2}\|\hat{\theta} - \theta^\star\|^2 + \|\hat{\theta} - \theta^\star\|\mu\sqrt{C(\mathcal{M})}(2\nu(f^\star) + \zeta)\sqrt{p}. \end{aligned}$$

A simple calculation then gives the first claim.

To prove the second claim of Theorem III.6, notice that from (III.10) we deduce that

$$\mathrm{Pers}(f_{\hat{\theta}}) \leq \mathrm{Pers}(f_{\theta^\star}) + \frac{1}{\mu}(P - P_n)(\gamma(\hat{\theta}, \cdot) - \gamma(\theta^\star, \cdot)) + \frac{1}{\mu}(P\gamma(\theta^\star, \cdot) - P\gamma(\hat{\theta}, \cdot)).$$

Since $\frac{1}{\mu}(P\gamma(\theta^\star, \cdot) - P\gamma(\hat{\theta}, \cdot)) \leq 0$, the claim immediately follows from the same arguments as in the first part (control of terms A and B, and reinjecting the control for $\|\hat{\theta} - \theta^\star\|$).

III.6.2 Proof of Theorem III.5

We want to use the results of [Bv11, Section 6] on the Lasso. The design matrix $\mathbf{X}$ here is given by $\mathbf{X}_{i,j} = \Phi_j(X_i)$, i.e. the i-th column of $\mathbf{X}$ is $\Phi(X_i)$, denoting as done earlier $\Phi(x) = (\Phi_1(x), \ldots, \Phi_p(x)) \in \mathbb{R}^p$. The empirical Gram matrix is

$$\hat{\Sigma} = \frac{1}{n}\mathbf{X}^T\mathbf{X} = \frac{1}{n}\sum_{i=1}^n \Phi(X_i)\Phi(X_i)^T.$$

Following the same argument as in the proof of Lemma III.10 leading up to (III.9), we have $\mathbb{E}(\hat{\Sigma}) = I_p$ and, according to Theorem 5.1.1 of [Tro15]:

$$\mathbb{P}(\Lambda_{\min}(\hat{\Sigma}) \leq 1/2) \leq p\left(\frac{e^{-1/2}}{\sqrt{1/2}}\right)^{\frac{n}{C(\mathcal{M})p}}.$$

(The minor difference in comparison to (III.9) is due to the fact that we only need the upper bound on the largest eigenvalue from Theorem 5.1.1 of [Tro15] here.) Therefore, we have

$$\mathbb{P}(\Lambda_{\min}(\hat{\Sigma}) \geq 1/2) \geq 1 - p\exp\left(\frac{n}{C(\mathcal{M})p}\ln\left(\frac{2e^{-1/2}}{\sqrt{2}}\right)\right) \geq 1 - \exp\left(-\frac{0.15n}{C(\mathcal{M})p} + \ln(p)\right)$$

This shows that, for $p \log p/n$ large, the smallest eigenvalue of the empirical Gram matrix is larger than $1/2$ with high probability, and the compatibility condition is verified with a constant larger than $1/2$.

We can now use the results of [Bv11, Section 6]. Their Theorem 6.1 remains true for the norm $I(\theta) = \sum_{i=1}^p |\theta_i| \mathrm{Pers}(\Phi_i)$ and a compatibility constant larger than $1/2$ with large probability. For a given μ_0, we therefore have, for $\mu \geq 2\mu_0$, with probability larger than $1 - \exp\left(-\frac{0.15n}{C(\mathcal{M})p} + \ln(p)\right)$,

$$\frac{1}{n}\|\mathbf{X}(\hat{\theta} - \theta^\star)\|_2^2 + \mu \sum_{i=1}^p \mathrm{Pers}(\Phi_i)|\hat{\theta}_i - \theta_i^\star| \leq \frac{\mu^2 s}{2},$$

on the event

$$\mathcal{E} := \left\{ \max_{1 \leq j \leq p} \frac{2}{n} \left| \sum_{i=1}^n \varepsilon_i \Phi_j(X_i) \right| \leq \mu_0 \right\}.$$

For a well chosen value μ_0, the event $\mathcal{E}$ is realized with large probability. Indeed, sub-Gaussianity of the ε_i's (with parameter σ^2) yields, by using similar arguments as given in the proof of Theorem III.6, that conditionally on the X_i's, the random variable $\sum_{i=1}^n \varepsilon_i \Phi_j(X_i)$ is sub-Gaussian with parameter $\sigma^2 \sum_{i=1}^n \Phi_j(X_i)^2$. We thus obtain

$$\mathbb{P}_\varepsilon \left(\frac{1}{n} \left| \sum_{i=1}^n \varepsilon_i \Phi_j(X_i) \right| > \mu_0 \right) \leq 2 \exp\left(\frac{-n^2 \mu_0^2}{4\sigma^2 \sum_{i=1}^n \Phi_j(X_i)^2} \right),$$

where we use the fact that for any sub-Gaussian random variable Y with parameter τ^2, we have $P(|Y| > \lambda) \leq 2\exp(-\lambda^2/4\tau^2)$ (e.g. see [Ver18]. Furthermore, Lemma III.8 gives that $\Phi_j(X_i)^2 \leq \sum_{j=1}^p \Phi_j(X_i)^2 \leq C(\mathcal{M})p$ almost surely, and thus we have

$$\mathbb{P}_\varepsilon \left(\frac{1}{n} \left| \sum_{i=1}^n \varepsilon_i \Phi_j(X_i) \right| > \mu_0 \right) \leq 2 \exp\left(\frac{-n\mu_0^2}{4\sigma^2 C(\mathcal{M})p} \right).$$

A simple union-bound and the choice

$$\mu_0 = 2\sigma \sqrt{\frac{pC(\mathcal{M})(\ln(p) + x)}{n}}$$

directly gives the required result.

IV Statistical learning on measures, an application to persistence diagrams

In the previous section, we have seen that persistence diagrams can be split between a noise and a signal component. By introducing an adequate penalty, we can eliminate the noisy part, resulting in a smooth estimate of a function. This approach is particularly amenable to a model where we observe a smooth function $f^\star$ and noisy observations of the form $y_i = f^\star(x_i) + \varepsilon_i$, where ε_i is a noise term. However, in many cases, data cannot be modelled in such a way. Especially when considering a classification problem, we are only interested in grouping the data into different classes, with little consideration of whether they have been corrupted by noise. On a high level, we assume that the data carry some topological information that we want to use as classification features. We, therefore, transform the data into persistence diagrams and consider the diagrams as our new input data. We are now interested in using all the information in the diagrams, not only the one away from the diagonal. As persistence diagrams can be seen as discrete measures on $\mathbb{R}^2$, we make this problem fit in a more general setting of measure classification. More precisely, we consider a binary supervised learning classification problem where we observe measures on a compact space $\mathcal{X}$ instead of having data in a finite-dimensional Euclidean space. Formally, we observe data $D_N = (\mu_1, Y_1), \dots, (\mu_N, Y_N)$, 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 is a multi-instance learning problem. However, our approach allows more flexibility and diversity in the input data we can deal with. Besides persistence diagrams, we will describe several cases of applications such as time series and flow cytometry. 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. The work from this section has been submitted in [HBL23] and is joint with Gilles Blanchard and Clément Levrard.

Contents

IV.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 [DLLP97] 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 [Amo13] for a comprehensive review of the methods used in multi-instance classification. Closer to our work is the paper by [ST12], which studies the properties of MIL from a statistical learning perspective. More general are the works on distribution regression initiated by [HB05] for learning on general metric spaces. For instance, [MFDS12] tackles the case of classification of distribution and [PSRW13] that of regression. The theory for simple kernel estimators has been developed in [SSPG16]. Another recent perspective regarding distribution learning follows the works by [MC20] and [KKCM23], 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 [CLR21], 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 [ST12], while we allow for more general inputs. We can therefore derive generalization error bounds, see [MRT18] for an introduction to these concepts. We introduce specific classification algorithms which fit into this framework and that discriminate 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 [DDQHD13], flow cytometry [AFH+13], time series (possibly with an embedding mapping them in some Euclidean space), and text classification using a word embedding method such as WORD2VEC, see [MCCD13]. 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 [EH22] for an overview of the construction of this object and of 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 coor-

dinates 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 [B+15], persistence images [AEK+17], or Atol [CLR21] immediately get rid of the measure representation and transform the data into a vector. It then becomes possible to plug these vector representations into a standard classifier, we refer to [OHK18] for classification using linear classifiers. Some papers use kernel methods, such as [CCO17] or [LY18], while some other works make use of neural networks, such as [CCI+20], and more recently [RCB21]. We refer to the survey [HMR21] 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 in the present work 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 [BHPW20] 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 [Owa22].

The rest of Section IV decomposes as follows: in Section IV.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 IV.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 IV.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 publicly available[4]. Finally, Section IV.5 is devoted to the proofs of all the theoretical results.

IV.2 Statistical learning on measures

IV.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 $\mathcal{X}$. 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 [FV06] 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 [Amo13] for a complete survey, and cover many useful applications, from which we can cite image

[4] https://github.com/OlympioH/BBA_measures_classification

classification based on a finite number of descriptors as done in [WYHY15], flow cytometry (see Section IV.4), or text classification where each word is represented by a point in a high-dimensional space. Closer to us is the work by [CLR21] 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.

IV.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 conditionally on a sample $(Z_1, \dots, Z_N)$ which we recall is defined by

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

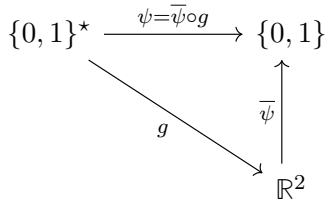
where $(\sigma_1, \dots, \sigma_N)$ are independent Rademacher random variables. The Gaussian complexity obeys the same definition where the σ_i are independent standard normal variables. The Rademacher complexity is a usual quantity in statistical learning that measures the richness of a set of functions. 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 of [SSBD14] for more details. It is common to upper bound this quantity 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 $\mathrm{VC}(\mathcal{F})$ the Vapnik-Chervonenkis dimension of a set of functions (or its pseudo-dimension in the case of real hypotheses classes) and by $\mathrm{VC}(\mathcal{F}, \gamma)$ its γ-fat shattering dimension. We refer to Chapter 6 of [SSBD14] 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.

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 [ST12]. In Proposition IV.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 $\mathcal{G}$. Assume the bag-function ψ to 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 $\sigma \in \mathfrak{S}_n$. Then there exist two functions g and $\overline{\psi}$ such that ψ decomposes as follows:



The function g is defined as $g(y_1, \dots, y_n) = (\sum_{i=1}^n y_i/n, n)$, i.e. it maps a sequence of zeros and ones to the proportion of ones 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 $h : \mu = \sum_{i=1}^n \delta_{x_i} \mapsto \psi(f(x_1), \dots, f(x_n))$, where $\psi \in \mathcal{G}$ and $f \in \mathcal{F}$.

Proposition IV.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 $\overline{\psi}$ is taken from a class $\overline{\mathcal{G}}$ 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 IV.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 $\mathcal{M}_m(\mathcal{X})$. We now propose to extend these results to the case of general measures with finite mass and therefore extend the MIL framework.

Generic measures, Lipschitz loss. In this section, using $\mathcal{Y} = \{-1, 1\}$ we build classifiers of the form $\operatorname{sgn}(g(\mu))$ for g in some function class $\mathcal{G}$ over $\mathcal{M}(\mathcal{X})$. Consider a κ-Lipschitz loss function $\mathcal{L}$. By the contraction principle for Rademacher complexities, it holds $\mathcal{R}_N(\mathcal{L} \circ \mathcal{G}) \leq \kappa \mathcal{R}_N(\mathcal{G})$. We therefore focus on the control of the Rademacher complexity of the class of real-valued predictors. We first extend Lemma 12 from [ST12] 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 p-norm, defined for two functions f_1 and f_2 in $\mathcal{F}$, and a sample of N measures $(\mu_1, \dots, \mu_N)$ as:

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

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

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

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

We defer the proof to Section IV.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 IV.3. *There exists an absolute constant K such that*

$$\mathcal{R}_N(\tilde{\mathcal{F}}) \leq \frac{K\overline{M}_2\sqrt{\mathrm{VC}(\mathcal{F})}}{\sqrt{N}}.$$

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 IV.4. *There exists an absolute constant K' such that*

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

The bounds from Theorems IV.3 and IV.4 match and are both of order $1/\sqrt{N}$, up to logarithmic 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 IV.5. 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 [MRT18] for instance) that the VC-dimension of Euclidean balls is $d+1$. We therefore have by Theorem IV.3 that there exist constants K and K' such that:

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

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 $\{-1, 1\}$. If the function ψ is fixed as it is the case in [ST12] 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 $h : \mu \mapsto \psi\left(\int_{\mathcal{X}} f(x)\mathrm{d}\mu(x)\right)$ 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 IV.6. *Assume that the class $\mathcal{G}$ consists of L-Lipschitz functions. Assume the null function $x \mapsto 0$ belongs to $\mathcal{F}$. Then there exist constants C_1 and C_2 such that for any sample of measures $\overline{\mu} = (\mu_1, \dots, \mu_N)$,*

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

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 IV.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 [Mau16] 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 IV.7. 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 IV.2.3. In this case, we directly have that $\mathbf{R}(\mathcal{G}) = \mathbb{E}[|\gamma|] = 1$. Therefore, keeping the same notation as above, in this scenario there exist universal constants C_1 and C_2 such that

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

IV.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 $\mu \mapsto \mathbb{1}\{\varepsilon\mu(A) - s \geq 0\}$.

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 discretized grid of parameters for a threshold $s \geq 0$ and for the set $\mathcal{A}$, and select those that minimize the empirical classification error:

$$(A^+, t^+) = \operatorname*{Arg\,Min}_{A,t} \mathcal{L}_+(A, t),$$

where

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

We similarly minimize the empirical classification error for reversed labels: $(A^-, t^-) = \operatorname{Arg\,Min}_{A,t} \mathcal{L}_-(A, t)$, for

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

If $\mathcal{L}_+(A^+, t^+) \leq \mathcal{L}_-(A^-, s^-)$ we set $\varepsilon = 1$ and pick (A^+, t^+), otherwise we set $\varepsilon = -1$, along with the corresponding set of parameters.

If all the measures μ_i write as a finite sum of Dirac measures, this step is very similar to MIL, since each of the N_i points in the bag μ_i will be assigned a label according to whether it belongs to the set A or not. The additional component 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 extends the theory developed in Chapter 3 of [ST12] about binary MIL where the aggregation function must be fixed. We therefore fit exactly within the framework of Proposition IV.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 $D_N = (\mu_i, Y_i)_{i=1}^N$:

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

where P 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 $P \circ f_{C,r,s,\sigma}$ 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

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

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 [ST12]. The class $\mathcal{F}$ is a smoothed version of ball indicators and has the same VC dimension: $\mathrm{VC}(\mathcal{F}) = d+1$. Using Proposition IV.6 with the class $\mathcal{G} = \{x \mapsto P(x-s)\}$, we can therefore write the corresponding generalization bound, using

that the cross-entropy loss is 1-Lipschitz. According to Theorem 26.5 of [SSBD14], we have that with probability at least $1-\delta$, for all $\theta = (C, r, \sigma, s) \in \mathcal{X} \times \mathbb{R}_+^2 \times [-S, S]$,

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

with universal constants C_1, C_2 and C_3.

Aggregation by boosting The two methods presented above 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 $\mathcal{X}$ but has obviously limited expressivity capabilities. We therefore propose in practice to combine several base "weak learners" with a boosting approach. We have implemented the ADABOOST method [FS+96], which classically calls the base method iteratively, giving more weight to misclassified data. In addition to greatly improving the predictive performance as opposed to selecting a single convex set, performing 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.

IV.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.

IV.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 Section I.1 for a presentation of the key concepts. For the sake of consistency, we recall the definition of the *Čech complex* and *Čech filtrations* built over point clouds as they are on of the main construction to build persistence diagrams.

Definition IV.8. Let $\mathbb{X} \subseteq \mathbb{R}^d$ be finite. The *Čech complex at scale* $r \geq 0$ is the simplicial complex $\check{\mathcal{C}}(\mathbb{X}, r)$ defined as follows: for $(x_0, \ldots, x_k) \in \mathbb{X}^{k+1}$, the simplex $\{x_0, \ldots, x_k\}$ is in $\check{\mathcal{C}}(\mathbb{X}, r)$ if the intersection of closed balls $\cap_{l=0}^k \overline{B}(x_l, t)$ is non-empty.

The key to persistence theory is to consider simplicial complexes with a multi-scale approach and consider a sequence of nested complexes rather than a single complex. To that extent, we can define a *filtration* of a simplicial complex as follows:

Definition IV.9. Consider a finite simplicial complex $\mathcal{K}$ and a non-decreasing function $f : \mathcal{K} \to \mathbb{R}$, in the sense that $f(\sigma) \leq f(\tau)$ whenever σ is a face of τ. We have that for every $a \in \mathbb{R}$, the sublevel set $\mathcal{K}(a) = f^{-1}(-\infty, a]$ is a simplicial subcomplex of $\mathcal{K}$. Considering all possible values of f leaves us with a nested family of subcomplexes

$$\varnothing = \mathcal{K}_0 \subset \mathcal{K}_1 \subset \ldots \subset \mathcal{K}_n = \mathcal{K},$$

called a *filtration*, where $a_0 = -\infty < a_1 < a_2 < \ldots < a_n$ are the values taken by f on the simplices of $\mathcal{K}$.

For instance, considering all possible scales of the Čech complex of a point set $\mathbb{X}$ naturally defines a filtration over the complete simplicial complex $\mathcal{K} = 2^{\mathbb{X}}$. The Čech filtration of $\mathbb{X}$, denoted by $\check{\mathcal{C}}(\mathbb{X})$ is equivalent to centering balls around each point of $\mathbb{X}$ and have the balls' radii grow from 0 to ∞. For general filtrations, as the scale parameter grows we are interested in tracking the evolution of the Betti numbers of the simplicial complexes. If a k-dimensional hole starts to exist at some time b and disappears at time d in the filtration, we add the point (b, d) in the k-dimensional persistence diagram D_k of the filtration. A persistence diagram 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: $\xi_k = \sum_{(b,d)\in D_k} \delta_{(b,d)}$ to conform to the theory and algorithms developed in Section IV.2.2.

We illustrate the construction of 0, 1 and 2-persistence diagrams of a Čech filtration in Figure 31 where we sample n points uniformly on a torus, according to an algorithm provided by [DHS+13]. 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 IV.12. Following this approach, works such as [AEK+17] 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 [BHPW20], or dimension. We give further evidence of that claim 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.

Finally, note that the success of persistence diagrams in topological data analysis has been motivated by the possibility to compare diagrams (with possibly different number of points) using distances inspired by optimal transport, the most popular being the *bottleneck distance*, which benefits from some stability properties, [CSEH07]:

Definition IV.10. Let $\Delta = \{(x, x)|x \in \mathbb{R}\}$ be the diagonal of $\mathbb{R}^2$.

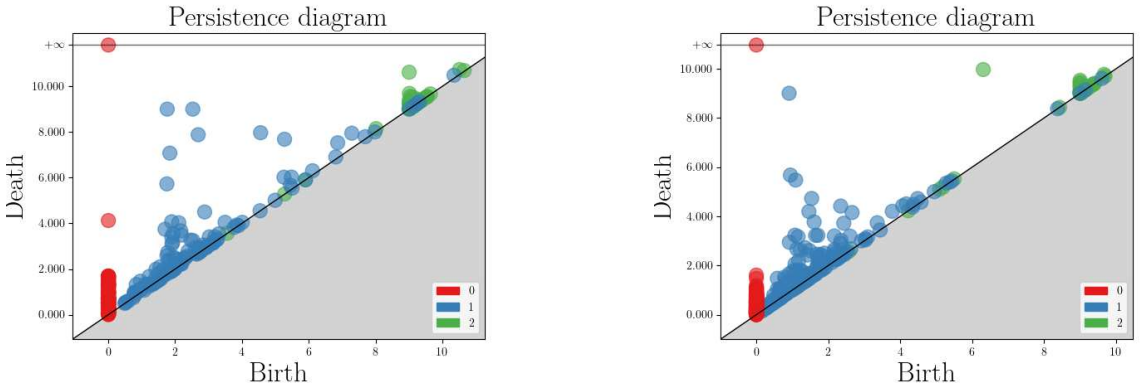
The *bottleneck distance* d_B between two persistence diagrams D and D' is defined by:

$$d_B(D, D') = \inf_{\eta: D\cup\Delta \to D'\cup\Delta} \sup_{x\in D\cup\Delta} \|x - \eta(x)\|_\infty,$$

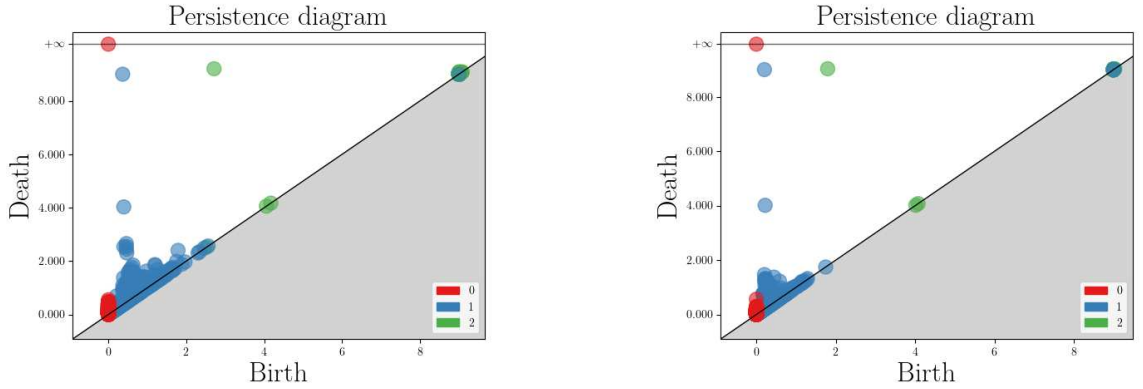
where the infimum is taken over all bijections η from $D \cup \Delta$ to $D' \cup \Delta$.

IV.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 Čech complex is the global one corresponding to the homology of the manifold supporting



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



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

Figure 31: 0, 1 and 2-persistence diagrams for n points uniformly sampled on a torus.

the data. 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 IV.11. 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

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

In the following theorem, dgm denotes the concatenation over all dimensions of the persistence diagrams of a filtration.

Theorem IV.12. *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 $K_i = \min_{(p,q)\in \mathrm{dgm}(\check{C}(\mathbb{M}_i))} |q - p|$ for $i = 1, 2$. Assume that there exists $K_3 > 0$ such that*

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

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 IV.5.6. We refer to [CDSO14] for the construction of the Čech filtration of possibly infinite metric spaces (and not simplicial complexes as previously), which ensures that $\mathrm{dgm}(\check{\mathrm{C}}(\mathbb{M}_1))$ and $\mathrm{dgm}(\check{\mathrm{C}}(\mathbb{M}_2))$ 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. Studying geometric quantities of the Čech complex of a random point cloud $\mathbb{X}_n = (X_i)_{i=1}^n$ on $\mathbb{R}^d$ is a deeply studied problem and we refer to [BA14] for some preliminary results regarding the critical points of the distance function to a point cloud. Some results have been adapted in [BM15] to point clouds sampled on manifolds. Finally, we refer to [BK18] for a detailed survey on random geometric complexes. Assume that we consider the Čech complex at a scale r_n that decreases with n and such that $r_n \to 0$. The speed at which r_n tends to 0 as $n \to \infty$ is paramount and dictates the type of results we can expect. In what follows, we focus on the *sparse regime*, i.e. $nr_n^d \to 0$ as $n \to \infty$. In this regime, asymptotic properties of the persistence diagram of the Čech filtration have been studied in [Owa22]. When considering persistent quantities, we consider the Čech complex at all possible ranges and we renormalize the sample points themselves by the sequence r_n. We generalize the results of the above-mentioned citation in the following theorem where our contribution is two-fold: the data are now allowed to be sampled from a manifold, and we provide a rate of convergence of the persistence diagram towards its limiting measure. Before stating the theorem, we define the function h_r by

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

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

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

Theorem IV.13. *Let* M *be a closed orientable* $\mathcal{C}^2$ *Riemannian manifold of dimension* d *with reach* $\tau_M \geq \tau_{\min}$. *Let* $\mathbb{X}_n = (X_i)_{i=1}^n$ *be an i.i.d. sample drawn from a* L*-Lipschitz density* f *on the manifold where* $\mathcal{H}$ *is the Hausdorff measure on the manifold.*

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*

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

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*

$$\xi_{k,n}(R_{s,t,u,v}) = \frac{Card(r_n R_{s,t,u,v} \cap dgm_k(\check{C}(\mathbb{X}_n/r_n))}{n^{k+2} r_n^{d(k+1)}},$$

which counts the number of points of the k-th persistence diagram of the rescaled data falling in the rectangle $r_n R_{s,t,u,v}$. Assume that we are in the sparse divergence regime, i.e. the sequence r_n verifies:

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

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.

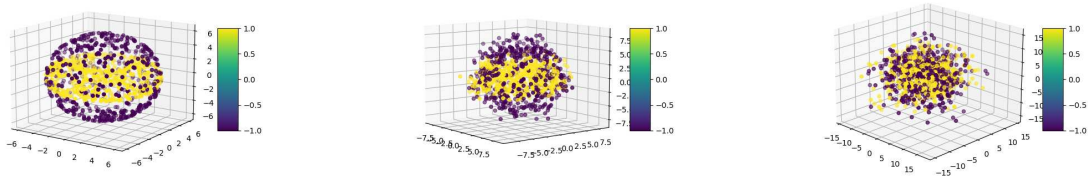
We defer the proof to Section IV.5.7.

This theorem asserts that asymptotically, the rescaled persistence diagram of the Čech filtration built on an adequately rescaled point cloud on $\mathbb{R}^d$ converges to a measure μ_k which depends on f only through $\int_M f^{k+2} \mathrm{d}\mathcal{H}$. Moreover, given two distributions f_1 and f_2 such that there exists $k \in [\![0, d-1]\!]$ such that $\int_M f_1^{k+2} \mathrm{d}\mathcal{H} \neq \int_M f_2^{k+2} \mathrm{d}\mathcal{H}$, 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 IV.14. *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 $\int_M f_1^{k+2} \mathrm{d}\mathcal{H} \neq \int_M f_2^{k+2} \mathrm{d}\mathcal{H}$. Let $0 < s \leq t \leq u \leq v$. For n large enough, the number of points in the persistence diagram falling in the rectangle $R_{s,t,u,v}$ identifies the correct sampling density 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.*

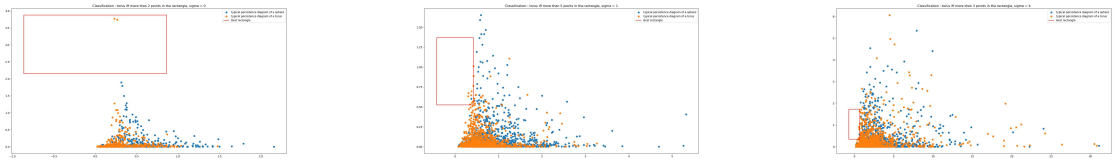
The proof of Corollary IV.14 is a straightforward consequence of the Chebyshev's inequality and we defer it to Section IV.5.8. Deriving a finer concentration inequality is still an open question: indeed, we have only used a bound on the variance of the random variable $\xi_{k,n}$ to use Chebyshev's inequality. 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 [YSA17] to our framework in the sparse regime.

The results derived in Proposition IV.12 and Corollary IV.14 both state that the number of sample 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 IV.1 and $\overline{M}_2$ in Theorem IV.3 and Proposition IV.6 assert that the number of points in the diagram (directly related to the number of sample points) must not be too large in order to obtain a good control of the Rademacher complexity. The number of sample points n acts as a trade-off between the separation of the two classes and a control of the predictive risk.



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

Figure 32: Data to classify. Yellow: torus, purple: sphere.



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

Figure 33: Best rectangle to classify points from a sphere or a torus.

IV.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 Čech complex. The persistence diagram of the torus is expected to have two high-persistence features. Some examples of data are shown in Figure 32 and rectangle classifiers on Figure 33. 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 IV.12. 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 34 shows the first four rectangles of the boosting procedure.

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

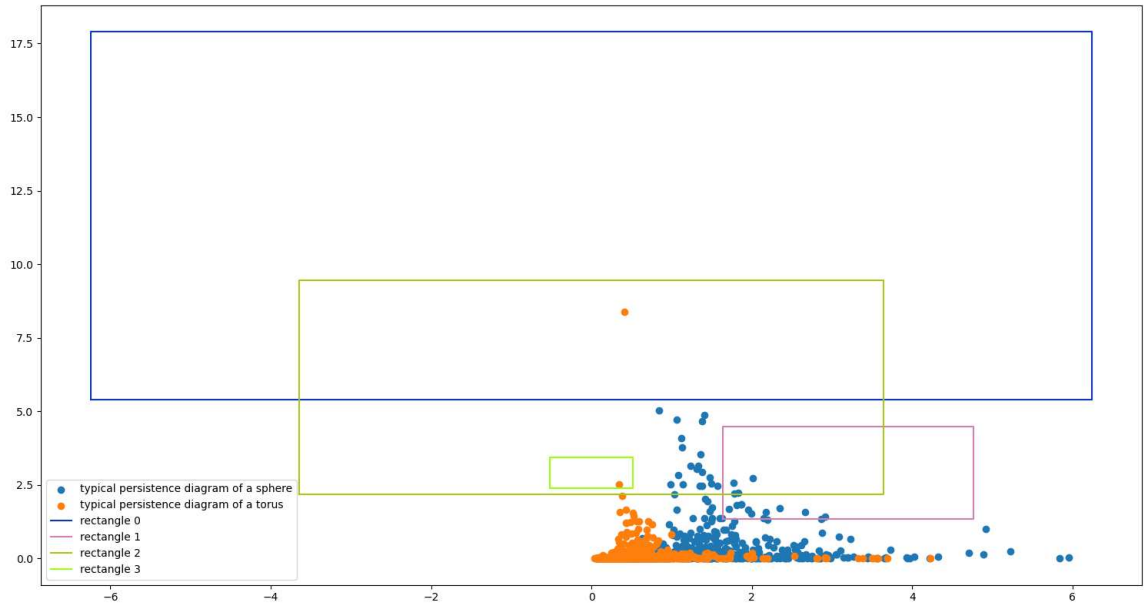
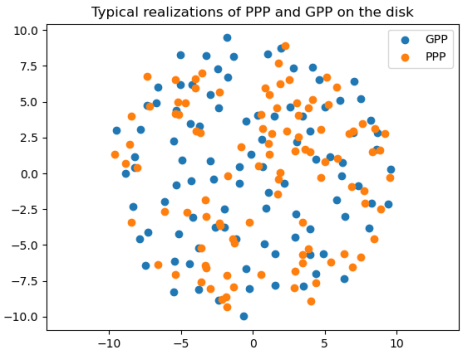


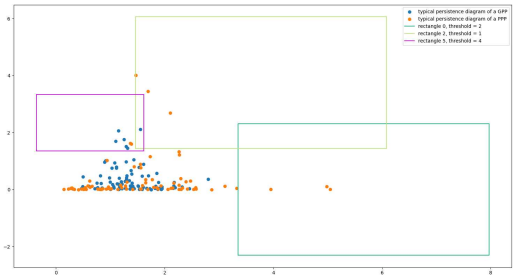
Figure 34: Boosting for manifold classification.

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 [OHK18]. 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 model has been trained on 400 processes and tested on 200. We have reached similar classification accuracy (around 94% in both cases). [OHK18] 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 IV.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 35a. 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 35b. 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).



(a) Samples of point processes



(b) Boosting rectangles

Figure 35: Point processes classification problem.

IV.4 Quantitative experiments

We 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 IV.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 IV.2.2 warrant for the validity of this sub-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.[5]

IV.4.1 Persistence diagrams

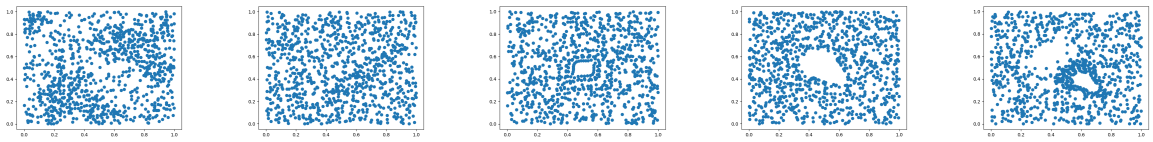
`ORBIT5K` dataset The dataset `ORBIT5K` is often used as a standard benchmark for classification methods in TDA. This dataset consists of subsets of size 1000 of the unit cube $[0,1]^2$ generated by a dynamical system that depends on a parameter $\rho > 0$. To generate a point cloud, a random initial 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 36; $\rho \in \{4.0, 4.1, 4.3\}$ 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 8, where the results are averaged over 10 runs. We compare our scores to four kernel methods on persistence diagrams taken respectively from [RHBK15], [KHF16], [CCO17], [LY18], and two

[5] https://github.com/OlympioH/BBA_measures_classification



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

Figure 36: 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 8: Classification scores for the ORBIT5K dataset.

methods that use a neural network architecture to vectorize persistence diagrams: [CCI+20] and [RCB21]. 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 done by [CCI+20], 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 9. 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 [CB20]). 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, [CCI+20] which is a neural network that learns a vector representation of a persistence diagrams and Atol, [RCL+21] which is another measure learning method. Note that on the biggest dataset COLLAB, our method is clearly outperformed by the other methods, especially Atol.

IV.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

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 9: Classification scores for graph data.

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 10: Classification scores for multi-dimensional time series dataset.

our model on the Acute Myeloid Leukemia (AML) dataset available here[6]. 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 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 [AFH+13]. 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 [BDL+18]. We compare our method against standard classification methods, and report the results from [BB21] in Table 10

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 $\mathbb{R}^d$.

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

IV.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 IV.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 11 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 logistic 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` ([MBGY14]) and `Scikit-learn` ([PVG+11]) 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 15. It is also to be noted from Table 11 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 12 where we also compare our running times with the Atol method from [CLR21] 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 section, 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 IV.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.

Table 11: Computational time (in seconds), torus versus 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 12: 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	-

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 15 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.

IV.5 Proofs

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

IV.5.1 Proof of Proposition IV.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 [SSBD14], we have that for every integer k, $\gamma_{\mathcal{H}}(k) \leq \gamma_{\tilde{\mathcal{F}}}(k)\gamma_{\overline{\mathcal{G}}}(k)$ using the observation above Proposition IV.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 [ST12] yields the wanted result.

IV.5.2 Proof of Lemma IV.2

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

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

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

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 :

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

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

IV.5.3 Proof of Theorem IV.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

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

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

$$
\begin{aligned}
\mathcal{R}_N(\tilde{\mathcal{F}}) &\leq \frac{12}{\sqrt{N}} \int_0^{\overline{M}_2} \sqrt{\ln \mathcal{N}(\mathcal{F}, L_2(\overline{\mu}), \varepsilon/\overline{M}_2)} d\varepsilon \text{ by the above lemma,} \\
&\leq \frac{12}{\sqrt{N}} \int_0^{\overline{M}_2} \sqrt{K_0 \mathrm{VC}(\mathcal{F}, c\varepsilon/\overline{M}_2) \ln(2\overline{M}_2/\varepsilon)} d\varepsilon \text{ by Theorem 1 of [MV03],} \\
&\leq \frac{12\overline{M}_2}{\sqrt{N}} \int_0^1 \sqrt{K_0 \mathrm{VC}(\mathcal{F}, c\varepsilon) \ln(2/\varepsilon)} d\varepsilon \text{by a change of variables,} \\
&\leq \frac{K_1 \overline{M}_2 \sqrt{\mathrm{VC}(\mathcal{F})}}{\sqrt{N}} \int_0^1 \sqrt{\ln(2/\varepsilon)} d\varepsilon.
\end{aligned}
$$

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

IV.5.4 Proof of Theorem IV.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

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

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:

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

In particular, by taking $x_1, \dots x_N$ that are $2\varepsilon/\overline{M}_2$-shattered by $\mathcal{F}$ if $N \leq \mathrm{VC}(\mathcal{F}, 2\varepsilon/\overline{M}_2)$ along with uniform weights, Proposition 1.4 from [Tal03] states that the logarithm of the packing number dominates the fat-shattering function. If $N > \mathrm{VC}(\mathcal{F}, 2\varepsilon/\overline{M}_2)$, the same result simply follows by considering the uniform measure on $\mathrm{VC}(\mathcal{F}, 2\varepsilon/\overline{M}_2)$ 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 $\varepsilon > 0$,

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

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

IV.5.5 Proof of Proposition IV.6

Let $\overline{\mu} = (\mu_1,\dots,\mu_N)$ be a sample of N measures. In Theorem 2 of [Mau16], 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 $f_0 \in \mathcal{F}$,

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

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

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 IV.3 permits to control the first term:

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

Analogously to the proof of Theorem IV.3, we can simply bound the diameter by $\mathrm{Diam}(\tilde{\mathcal{F}}(\overline{\mu})) \leq \sqrt{N\overline{M}_2}$, 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.

IV.5.6 Proof of Theorem IV.12

Assume that we observe a n-sample from $\mathbb{M}_1$. Corollary 3 of [CGLM14] states that for every $\varepsilon > 0$,

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

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

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

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

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

By assumption, for $i = 1, 2$, $\mathrm{dgm}(\check{\mathrm{C}}(\mathbb{M}_i))$ has no point at a distance less than K from the diagonal. We can now distinguish two cases :

- $\mathrm{dgm}(\check{\mathrm{C}}(\mathbb{M}_1))$ and $\mathrm{dgm}(\check{\mathrm{C}}(\mathbb{M}_2))$ have the same number of points m (all these points are at least away from K to the diagonal). Under these circumstances, $\mathrm{dgm}(\check{\mathrm{C}}(\hat{X}_n))$ 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 (IV.1). Consider squares of size $K/2$ centered on the points of $\mathrm{dgm}(\check{\mathrm{C}}(\mathbb{M}_1))$ and $\mathrm{dgm}(\check{\mathrm{C}}(\mathbb{M}_2))$. If they all contain the same number of points from $\mathrm{dgm}(\check{\mathrm{C}}(\hat{X}_n))$, we have a contradiction with (IV.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 (IV.1), $\mathrm{dgm}(\check{\mathrm{C}}(\hat{X}_n))$ must have the same number of points as $\mathrm{dgm}(\check{\mathrm{C}}(\mathbb{M}_1))$ above $K/2$ and do not have the same number of points as $\mathrm{dgm}(\check{\mathrm{C}}(\mathbb{M}_2))$. Counting the number of points in the (infinite but truncatable) rectangle $\{(p,q) || q-p| > K/2\}$ is therefore enough to classify between the two metric spaces.

IV.5.7 Proof of Theorem IV.13

We define the persistent Betti number $\beta_{k,n}(a,b)$ as the number of k-holes of $\check{\mathrm{C}}(r_n^{-1}\mathbb{X}_n, r)$ that persist between $r=a$ and $r=b$. It corresponds to the number of points in the persistence diagram that falls in the upper-left quadrant having an angle at the point (a,b).

First note that $\mathrm{Card}(r_n R_{s,t,u,v}) = \beta_{k,n}(t,u) - \beta_{k,n}(t,v) - \beta_{k,n}(s,u) + \beta_{k,n}(s,v)$. As in [Owa22], we denote by

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

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 [Owa22, Lemma 4.1]

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

where

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

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

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

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 $(x_2, \dots x_{k+1}) \in \mathcal{B}(x_1, r_n(k+2)t^+)^k$ (recall that $t \le t^+$). Denoting by $\tilde{M}_n = h_n(M)$, with $h_n : u \mapsto \frac{u-x_1}{r_n}$, and using [Fed59, Theorem 3.1] leads to the change of variable

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

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 $\tilde{\tau} \ge 4(k+2)t^+$, see, for instance [AL19, Lemma 1]). Using [Fed59, Theorem 3.1] again yields for the change of variable $y_j = \exp_0(v_j)$ that

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

According to [AL19, Lemma 1], whenever $y_j \in \mathcal{B}_{\tilde{M}_n}(0, (k+2)t^+)$, we have $\|y_j - v_j\| \le C((k+2)t^+)^2 r_n/\tau_{\min}$, and $\|\,\mathrm{d}_{v_i}\exp_0 - I_d\|_{op} \le \frac{5}{4\tilde{\tau}_n} = \frac{5r_n}{4\tau} \le \frac{5r_n}{4\tau_{\min}}$, so that

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

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

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

Denoting by

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

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

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

We deduce that

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

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

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

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

according to [AL19, 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., [Lee90, Theorem 3]) yields that

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

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

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

so that

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

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

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

Plugging these inequalities into (IV.4) leads to

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

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

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

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

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

with $m = k+3$, so that

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

for n large enough.

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

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

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,

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

IV.5.8 Proof of Corollary IV.14

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

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

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 IV.13 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.

V Euler characteristic tools for topological data analysis

In the previous section, we have seen that we can transform raw data into persistence diagrams to classify them using underlying topological information. This approach has been particularly fruitful in the case of point cloud and graph data. We have developed algorithms and corresponding theoretical guarantees that work directly on the diagrams seen as discrete measures. We have demonstrated that this approach has a performance comparable to simple vectorizations of persistence diagrams (e.g. persistence images) coupled with a standard off-the-shelf classification algorithm (e.g. regularized logit). In addition, the computation time of our method is comparable to that of this standard method; see Table 11. When considering the complete pipeline that maps the raw data to a binary prediction, the computation of persistence diagrams occupies a fair share of the total computation time. We wonder whether some descriptors bypass the diagram computation step while carrying relevant and interpretable topological information. This section aims at providing an affirmative answer to this question. Indeed, we construct descriptors by analogy with the Euler characteristic of a simplicial complex, which carries information about the Betti numbers simply by counting vertices and without any homology computation. More generally, we study Euler characteristic techniques in topological data analysis. Pointwise computing the Euler characteristic of a family of simplicial complexes built from data gives rise to the Euler characteristic profile. We show that this simple descriptor achieves state-of-the-art performance in supervised tasks at a modest computational cost. Inspired by signal analysis, we compute hybrid transforms of Euler characteristic profiles. These integral transforms mix Euler characteristic techniques with Lebesgue integration to provide highly efficient compressors of topological signals. As a consequence, they show remarkable performances in unsupervised settings. Most notably, Euler characteristic profiles and hybrid transforms bypass the computation of persistence diagrams resulting in a substantial improvement in computational complexity while maintaining competitive performance. In addition, these descriptors naturally generalize to *multi-persistence*. Finally, we prove stability results for these descriptors and asymptotic guarantees in random settings.

Contents

V.1 Introduction

We have seen in the previous sections that persistence diagrams are a summary of the topological information contained in a multi-scale filtration of the data. The space of persistence diagrams is a metric space for the so-called *bottleneck distance*, [CSEH07], but it cannot be isometrically embedded into a Hilbert space [CB18, BW20]. Although it is possible to perform some machine learning tasks directly on the space of persistence diagrams seen as measure, as done in Section IV, most methods consist in transforming persistence diagrams into vectors. Most commonly used techniques include persistence images [AEK+17], landscapes [B+15], and more recently measure-oriented vectorizations in [RCL+21] and neural network methods from [CCI+20, RCB21]. An overview of topological methods in machine learning has been presented in the survey of [HMR21]. These methods have demonstrated their efficiency in a wide variety of applications and types of data, such as health applications [RYB+20, FM22, ACC+21], biology [IOH20, RB19] or material sciences [LBD+17, HNH+16].

In many practical scenarios, it is natural to look at data with more than one parameter, i.e., to consider multi-parameter families of topological spaces instead of one-parameter ones. It allows one to cope with outliers by filtering the space with respect to an estimated local density, or to deal with intrinsically multi-parameter data, such as blood cells with several biomarkers. However, there does not exist a complete combinatorial descriptor similar to the persistence diagram that could make them usable in artificial intelligence [CZ09]. One of the main objectives of this field is to build informative descriptors of such families. Although not intrinsically multi-parameter, persistence landscapes have successfully been generalized to the multi-parameter setting in [Vip20] and persistence images to the two-parameter setting in [CB20]. Besides their high level of sophistication, the main limitation of these tools is their computational cost; see [CB20, Table 2] and Section V.4.5.

In contrast, some topological methods do not compute homological information—thus bypassing the computation of persistence diagrams—but rather compute the Euler characteristic of the topological spaces at hand. The Euler characteristic of a simplicial complex is a celebrated topological invariant that is simply the alternated sum of the number of simplices of each dimension. Considering the pointwise Euler characteristic of a one-parameter family of simplicial complexes gives rise to a functional multi-scale descriptor called the *Euler characteristic curve*.

Though Euler characteristic based descriptors may appear coarse, we highlight four main reasons to favour them. First, they have a good predictive power [SZ21, JKN20, AQO+22]. Second, the simplicity of these descriptors translates into a reduced computational cost. They

can be computed in linear time in the number of simplices instead of typically matrix multiplication time for persistence diagrams [MMS11]. Moreover, the locality of the Euler characteristic can be exploited to design highly efficient algorithms computing Euler curves, as in [HW17]. Third, there are several known theoretical results on the Euler characteristic of a random complex. Mean formulae for the Euler characteristic of superlevel sets of random fields are proven in [AT09], and the limiting behaviour of the Euler characteristic of a complex built on a Poisson process are established in Corollary 4.2 of [BA14]. Furthermore, Euler curves associated with random point clouds are proven to be asymptotically normal for a well-chosen sampling regime in [KRP21], where the authors also apply this construction to bootstrap. Fourth, they naturally generalise to the multi-parameter setting, becoming so-called *Euler characteristic surfaces* [BSA+22] and *profiles* [DG22].

We demonstrate that these tools reach state-of-the-art performance at a minimal computational cost when coupled with a powerful classifier such as an XGB or a random forest. However, due to their simplicity, these descriptors do not manage to linearly separate the different classes or be competitive on unsupervised tasks. Inspired by signal analysis, we cope with these limitations by studying integral transforms of Euler characteristic curves and profiles. More precisely, we consider a general notion of integral transforms mixing Lebesgue integration and Euler characteristic techniques recently introduced in [Leb22] under the name of *hybrid transforms*. In the one-parameter case, hybrid transforms are classical integral transforms of Euler curves. Similarly, hybrid transforms depend on a choice of kernel which offers a wide variety of possible signal decompositions. Yet, hybrid transforms differ from classical integral transforms in general. In so doing, they enjoy many specific appealing properties, such as compatibility with topological operations from Euler calculus [Leb22, Section 5]. Most importantly, in the context of multi-parameter sublevel-sets persistence, hybrid transforms can be expressed as one-parameter hybrid transforms of Euler curves associated with a linear combination of the filtration functions. As a consequence, mean formulae for hybrid transforms associated with Gaussian random fields are derived in [Leb22, Section 8], and we prove here a law of large numbers in a multi-filtration set-up. Studying the asymptotic behaviour of topological descriptors of random complexes is a deeply-studied question in the one-parameter setting; see [BK18] for a survey. Together with the works of [BH22], our results form the first occurrence of limiting theorems in a multi-persistence framework in the literature.

Contributions and outline. After introducing the necessary notions in Section V.2, we provide heuristics on how to choose the kernel of hybrid transforms and give many examples of the type of topological and geometric behaviour Euler curves and their integral transform can capture from data in Section V.3. Most importantly, our main contributions are the following:

- We demonstrate that Euler profiles achieve state-of-the-art accuracy in supervised classification and regression tasks when coupled with a random forest or an XGB (Sections V.4.1, V.4.2 and V.4.4) at a very low computational cost (Section V.4.5). Note that the multi-parameter nature of our tools and their computational simplicity allows us to use up to 5-parameter filtrations to classify graph data.
- We demonstrate that hybrid transforms act as highly efficient information compressors. As a consequence, they outperform Euler profiles in unsupervised classification tasks and in supervised tasks when plugging a linear classifier (Figure 45 and Sections V.4.1 to V.4.3). In Section V.4.3, we illustrate their ability to capture fine-grained information on a real-world data set.

- We provide several theoretical guarantees for these descriptors. First, we prove stability properties that clarify the robustness of our tools with respect to perturbations (Section V.5). Expressed in terms of L_1 norms, these are also hints of the sensitivity of our tools to the underlying geometry of the data at hand. Then, we establish the pointwise convergence of hybrid transforms associated with random samples and their asymptotic normality for a specific filtration function. We also establish a law of large numbers in a multi-filtration set-up (Section V.6).

Finally, Section V.7 is devoted to the proofs of the results stated in Sections V.5 and V.6.

V.2 Definitions

In this section, we present all the notions used throughout this section. Let us first introduce some conventions.

(i) The dual of a vector space $\mathbb{V}$ is denoted by $\mathbb{V}^*$, and $\mathbb{R}^m$ will always be identified with its dual under the canonical isomorphism. For $\xi \in \mathbb{R}^{m*}$ and $t \in \mathbb{R}^m$, we often denote $\xi \cdot t = \xi(t)$.
(ii) We denote by $\mathbb{R}^{m*}_+$ the cone of linear forms on $\mathbb{R}^m$ that are non-decreasing with respect to the coordinatewise order on $\mathbb{R}^m$, or equivalently that have non-negative canonical coordinates.
(iii) Let I be an interval of $\mathbb{R}$ and denote by $L^1(I)$ the space of absolutely integrable complex-valued functions on I.
(iv) Let $p \in [1, \infty]$ and let $f : \mathbb{R}^m \to \mathbb{R}$ be locally p-integrable. We denote by $\|f\|_{p,M}$ the p-norm of $f \cdot \mathbf{1}_{[-M,M]^m}$. If f is p-integrable, we denote its p-norm by $\|f\|_p$.
(v) We always consider the coordinatewise order on $\mathbb{R}^m$.

V.2.1 Simplicial complexes, filtrations

For the sake of completeness, we hereby recall some notions about simplicial complexes and filtrations. A *(finite) abstract simplicial complex* $\mathcal{K}$, or simply *simplicial complex*, is a finite collection of finite sets that is closed under taking subsets. An element $\sigma \in \mathcal{K}$ is called a *simplex*, and subsets of σ are called *faces* of σ. The inclusion between simplices induces a partial order on $\mathcal{K}$ that we denote simply by $\leq$. The *dimension* of a simplex with k elements is equal to $k-1$. The *Euler characteristic* of a simplicial complex $\mathcal{K}$ is the integer:

$$\chi(\mathcal{K}) = \sum_{\sigma \in \mathcal{K}} (-1)^{\dim \sigma}.$$

Until the end of this section, we let $\mathcal{K}$ be a finite simplicial complex. An *m-parameter filtration* of $\mathcal{K}$ is a family $\mathcal{F} = (\mathcal{F}_t)_{t \in \mathbb{R}^m}$ of subcomplexes $\mathcal{F}_t \subseteq \mathcal{K}$ that is increasing with respect to inclusions, i.e., such that $\mathcal{F}_t \subseteq \mathcal{F}_{t'}$ for any $t, t' \in \mathbb{R}^m$ with $t \leq t'$. From now on, we do not refer explicitly to $\mathcal{K}$ when it is clear from the context. Many filtrations can be introduced by considering sublevel sets of functions:

Example V.1. Let $f : \mathcal{K} \to \mathbb{R}^m$ be a non-decreasing map, i.e., such that $f(\sigma) \leq f(\tau)$ for any $\sigma \leq \tau \in \mathcal{K}$. The map f induces an m-parameter filtration of $\mathcal{K}$ called *sublevel-sets filtration*, denoted by $\mathcal{F}_f$, and formed by the subcomplexes $(\mathcal{F}_f)_t = \{f \leq t\} := \{\sigma \in \mathcal{K} : f(\sigma) \leq t\}$ for any $t \in \mathbb{R}^m$. We sometimes refer to the function f as the *filter* of $\mathcal{F}_f$.

A lot of information on the geometry of a *point cloud*, that is, a finite subset of $\mathbb{R}^d$, is captured by its Čech complex:

Example V.2. Let $\mathbb{X} \subseteq \mathbb{R}^d$ be finite. The *Čech complex at scale* $t \geq 0$ is the simplicial complex $\check{\mathcal{C}}(\mathbb{X}, t)$ defined as follows. For $(x_0, \ldots, x_k) \in \mathbb{X}^{k+1}$, the simplex $\{x_0, \ldots, x_k\}$ is in $\check{\mathcal{C}}(\mathbb{X}, t)$ if the intersection of closed balls $\cap_{l=0}^{k} \overline{B}(x_l, t)$ is non-empty. The *filtered Čech complex*, or *Čech filtration*, is defined at each $t \in \mathbb{R}$ as the Čech complex at scale t for $t \geq 0$, and as the empty set for $t < 0$. For computational reasons, we rather use a homotopy equivalent complex in numerical experiments, called the *filtered alpha complex*, or *alpha filtration*, which is a subcomplex of the Delaunay triangulation; see [BE17].

The properties of the Čech complex of a random point cloud have been deeply studied theoretically. We refer to [BK18] and [Owa22] for the most recent results. When doing multi-parameter persistence, a common technique is to couple the Čech complex with some function on the data. Typically, we cope with outliers by coupling a Čech filtration with a density estimator built from the data at hand. This falls under the framework of function-Čech filtrations:

Example V.3. Let $\mathbb{X} \subseteq \mathbb{R}^d$ be finite and $f = (f_1, \ldots, f_m) : \mathbb{X} \to \mathbb{R}^m$ be a bounded function. The *function-Čech filtration* is the $(m+1)$-parameter filtration $\check{\mathcal{C}}(\mathbb{X}, f)$ of $2^{\mathbb{X}}$ defined for $r \in \mathbb{R}$ and $t = (t_1, \ldots, t_m) \in \mathbb{R}^m$ by:

$$\check{\mathcal{C}}(\mathbb{X}, f)_{(r,t)} = \left\{\sigma \in \check{\mathcal{C}}(\mathbb{X}, r) : \sigma \subseteq f_i^{-1}(-\infty, t_i],\ 1 \leq i \leq m\right\}.$$

Again, we rather use *function-alpha filtration* in numerical experiments, which are defined similarly using alpha complexes.

Let $\mathcal{F}$ be an m-parameter filtration and $\sigma \in \mathcal{K}$. The *support of* σ is the set $\mathrm{supp}(\sigma) := \{t \in \mathbb{R}^m : \sigma \in \mathcal{F}_t\}$. A filtration is called *finitely generated* if the support of any simplex appearing in the filtration is either empty or has a finite number of minimal elements; see Figure 37a for an illustration. Moreover, if the support of any simplex has at most one minimal element, then the filtration is called *one-critical*. In that case, we denote by $t(\sigma)$ the minimal element of $\mathrm{supp}(\sigma)$. For instance, function-Cech and function-alpha filtrations are one-critical. On the contrary, the degree-Rips bifiltration is not [LW16]. Note that sublevel-sets filtrations are one-critical. Conversely, any one-critical filtration is a sublevel set filtration for the function $f : \sigma \in \mathcal{K} \mapsto t(\sigma)$.

V.2.2 Euler characteristic tools

In this section, we recall the definitions of the descriptors of filtered simplicial complexes we use to perform topological data analysis. These invariants are defined using Euler characteristic profiles [BSA+22, DG22] and topological and hybrid transforms of constructible functions [Sch95, GR11, Leb22]. While these tools can be defined in the more general setting of o-minimal geometry, we focus on filtered simplicial complexes.

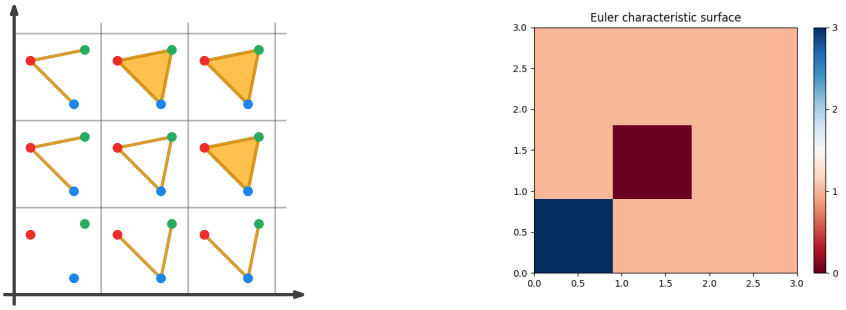
Given an m-parameter filtration, computing the Euler characteristic for every value of the parameter $t \in \mathbb{R}^m$ gives an integer-valued function on $\mathbb{R}^m$ that is a multi-scale descriptor of the evolution of the filtration with respect to t.

Definition V.4. The *Euler characteristic profile* of an m-parameter filtration $\mathcal{F}$ is the map:

$$\chi_{\mathcal{F}} : t \in \mathbb{R}^m \mapsto \chi(\mathcal{F}_t).$$

The map $\chi_{\mathcal{F}}$ is usually refered to as the *Euler characteristic curve* (ECC) of $\mathcal{F}$ when $m = 1$ and as the *Euler characteristic surface* (ECS) of $\mathcal{F}$ when $m = 2$; see [BSA+22, DG22].

We show in Figure 37 an Euler characteristic surface computed on an elementary example. Widely used in data analysis [SZ21, DG22, BSA+22, JKN20], this simple descriptor has proven to be efficient to capture meaningful information on the data at hand. However, as illustrated in the following sections, we are interested in more robust descriptors built from integral transformations.



(a) $\mathcal{F}$ (b) $\chi_{\mathcal{F}}$

Figure 37: A finitely generated 2-parameter filtration (a) and its associated Euler characteristic surface (b). All vertices have one birth time, while all other simplices have two.

Before introducing the other descriptors considered, we define the pushforward operation from Euler calculus; see [Sch89, Vir88]:

Definition V.5. Let $\mathcal{F}$ be a one-critical m-parameter filtration and $\xi \in \mathbb{R}_+^{m*}$. The *pushforward of* $\mathcal{F}$ *along* ξ is the one-parameter family defined for any $s \in \mathbb{R}$ by:

$$\left(\xi_*\mathcal{F}\right)_s = \bigcup_{\xi \cdot t \leq s} \mathcal{F}_t.$$

The *pushforward of* $\chi_{\mathcal{F}}$ *along* ξ is the Euler characteristic curve of $\xi_*\mathcal{F}$. We denote this curve by $\xi_*\chi_{\mathcal{F}}$. In other words, we have $\xi_*\chi_{\mathcal{F}} = \chi_{\xi_*\mathcal{F}}$. Writing the one-critical filtration as a sublevel set filtration, the pushforward operation has a simple expression:

Example V.6. Let $f : \mathcal{K} \to \mathbb{R}^m$ be a non-decreasing map and $\xi \in \mathbb{R}_+^{m*}$. The Euler characteristic profile of $\mathcal{F}_f$ is denoted by χ_f. It is an easy exercise to check that $\xi_*\mathcal{F}_f = \mathcal{F}_{\xi \circ f}$ and $\xi_*\chi_f = \chi_{\xi \circ f}$.

Introduced by [Sch95], the Radon transform plays a central role in topological data analysis. For instance, it allows one to prove that the so-called Euler characteristic transform and persistent homology transform are injective [GLM18, CMT22]. However, it has never been studied in data analysis as a topological descriptor of filtered simplicial complexes. To do so, we give its expression in this context.

Definition V.7. Let $\mathcal{F}$ be a one-critical m-parameter filtration. The *Radon transform* of $\chi_{\mathcal{F}}$ is the map:

$$\mathcal{R}_{\mathcal{F}} : (\xi, s) \in \mathbb{R}_+^{m*} \times \mathbb{R} \longmapsto \xi_*\chi_{\mathcal{F}}(s).$$

Hybrid transforms mixing Euler calculus and classical Lebesgue integration have been introduced in [Leb22]. In contrast to the Radon transform, these transforms are not purely topological. As a consequence, they are regular (continuous and piecewise smooth) and enjoy several beneficial properties, such as index theoretic formulae in the context of sublevel set persistence; see Propositions 4.1 and 4.2 and Theorem 8.3 in loc. cit.. In the present context, they can be defined as follows:

Definition V.8. Let $\mathcal{F}$ be a one-critical m-parameter filtration and $\kappa \in L^1(\mathbb{R})$. The *hybrid transform with kernel* κ of $\chi_{\mathcal{F}}$ is the map:

$$\psi_{\mathcal{F}}^{\kappa} : \xi \in \mathbb{R}_+^{m*} \mapsto \int_{\mathbb{R}} \kappa(s) \xi_* \chi_{\mathcal{F}}(s) \, \mathrm{d}s.$$

The following lemma is an obvious consequence of Example V.6. It states that any m-parameter hybrid transform restricted to an open half-line can be expressed as a one-parameter hybrid transform. It will be key to the proof of a law of large numbers for m-parameter hybrid transforms (Theorem V.18).

Lemma V.9. *Let $\mathcal{F}$ be a one-critical m-parameter filtration, let $\kappa \in L^1(\mathbb{R})$ and $\xi \in \mathbb{R}_+^{m*}$. For any $\lambda > 0$, one has:*

$$\psi_{\mathcal{F}}^{\kappa}(\lambda \xi) = \psi_{\xi_* \mathcal{F}}^{\kappa}(\lambda).$$

Euler characteristic profiles, Radon transforms, and hybrid transforms constitute the three descriptors of data we will use to perform topological data analysis. We give explicit expressions of these descriptors in two specific cases below. These formulae will allow us to design algorithms to compute them in Section V.3.1 and to build intuition on the type of behaviour they capture.

Connection with persistence diagrams. Suppose that $\mathcal{F}$ is a one-parameter filtration. Denote the corresponding k-th persistent diagram by $D_k = \left\{ \left(a_i^k, b_i^k\right) \right\}_{i=1,\dots,n_k}$ for real numbers $-\infty < a_i^k < b_i^k \leq \infty$ and an integer $n_k \geq 0$. It is then straightforward to check that:

$$\chi_{\mathcal{F}} = \sum_{k \geq 0} \sum_{i=1}^{n_k} (-1)^k \mathbf{1}_{[a_i^k, b_i^k)}. \tag{V.1}$$

Therefore, the Radon transform of $\chi_{\mathcal{F}}$ is:

$$\mathcal{R}_{\mathcal{F}} : (\xi, s) \in \mathbb{R}_+^* \times \mathbb{R} \mapsto \sum_{k \geq 0} \sum_{i=1}^{n_k} (-1)^k \mathbf{1}_{[\xi \cdot a_i^k, \xi \cdot b_i^k)}(s). \tag{V.2}$$

Let $\kappa \in L^1(\mathbb{R})$ and consider a primitive $\overline{\kappa}$ of κ. The hybrid transform with kernel κ of $\chi_{\mathcal{F}}$ is:

$$\psi_{\mathcal{F}}^{\kappa} : \xi \in \mathbb{R}_+^* \mapsto \sum_{k \geq 0} \sum_{i=1}^{n_k} (-1)^k \Big(\overline{\kappa}[\xi \cdot b_i^k] - \overline{\kappa}[\xi \cdot a_i^k] \Big), \tag{V.3}$$

with the convention that $\overline{\kappa}(\xi \cdot b_i^k)$ is the limit of $\overline{\kappa}$ at $+\infty$ when $b_i^k = +\infty$.

One-critical filtrations. Up to reducing $\mathcal{K}$, one can assume that for any $\sigma \in \mathcal{K}$, there is $t \in \mathbb{R}^m$ with $\sigma \in \mathcal{F}_t$. Then, one has:

$$\chi_{\mathcal{F}} = \sum_{\sigma \in \mathcal{K}} (-1)^{\dim \sigma} \mathbf{1}_{Q_{t(\sigma)}}, \tag{V.4}$$

where $Q_u := \{ t \in \mathbb{R}^m : t \geq u \}$ for any $u \in \mathbb{R}^m$. As a consequence, one has:

$$\mathcal{R}_{\mathcal{F}} : (\xi, s) \in \mathbb{R}_+^{m*} \times \mathbb{R} \mapsto \sum_{\sigma \in \mathcal{K}} (-1)^{\dim \sigma} \mathbf{1}_{[\xi \cdot t(\sigma), +\infty)}(s). \tag{V.5}$$

$\mathrm{L}\kappa \in L^1(\mathbb{R})$. Denote by $\overline{\kappa}$ the primitive of κ whose limit at $+\infty$ is 0. The hybrid transform with kernel κ of $\chi_{\mathcal{F}}$ is:

$$\psi_{\mathcal{F}}^{\kappa} : \xi \in \mathbb{R}_+^{m*} \longmapsto -\sum_{\sigma \in \mathcal{K}} (-1)^{\dim \sigma} \overline{\kappa}[\xi \cdot t(\sigma)]. \tag{V.6}$$

Remark V.10. We often define hybrid transforms by specifying the primitive $\overline{\kappa}$ of the kernel κ whose limit at $+\infty$ is 0. We call $\overline{\kappa}$ the *primitive kernel* of the hybrid transform.

Finally, in the case of a one-parameter filtration, hybrid transforms naturally appear as classical integral transforms of the Euler curve, making them a natural tool to extract information from the Euler curve and compress it into a small number of relevant coefficients.

Connection with classical transforms. Let $\mathcal{F}$ be a one-critical m-parameter filtration. First, assume that $m = 1$. For any $\xi \in \mathbb{R}_+^*$ and any $s \in \mathbb{R}$, one has $(\xi_* \mathcal{F})_s = \mathcal{F}_{s/\xi}$ and hence $\xi_* \chi_{\mathcal{F}}(s) = \chi_{\mathcal{F}}(s/\xi)$. A change of variables then ensures that the hybrid transform with kernel $\kappa \in L^1(\mathbb{R})$ is nothing but the rescaled classical transform:

$$\psi_{\mathcal{F}}^{\kappa} : \xi \in \mathbb{R}_+^* \mapsto \xi \cdot \int_{\mathbb{R}} \kappa(\xi \cdot s) \chi_{\mathcal{F}}(s) \, \mathrm{d}s. \tag{V.7}$$

Assume now that $m \geq 2$. The hybrid transform with kernel κ differs from the classical integral transform:

$$\xi \in \mathbb{R}_+^{m*} \mapsto \int_{\mathbb{R}^m} \kappa(\xi \cdot x) \chi_{\mathcal{F}}(x) \, \mathrm{d}x.$$

See [Leb22, Example 3.18] for a counter-example. In some special cases, however, such as when $\kappa(t) = \exp(-t)$, hybrid transforms and classical transforms coincide up to a rescaling [Leb22, Examples 5.12 and 5.17]. The interest of hybrid transforms over classical transforms can be motivated by the following example:

Example V.11. The one-parameter hybrid transform with kernel $\kappa(t) = \exp(-t)$ is also known as the *persistent magnitude* [GH21]. As proven in loc. cit. following the work of [Ott22], this object is related to another invariant of finite metric spaces called *magnitude* and introduced in [Lei13]. The formulation of persistent magnitude as a hybrid transform naturally generalizes it to the multi-parameter setting while preserving its most appealing properties [Leb22].

V.3 Method

In this section, we begin by describing the algorithms used to compute our descriptors as well as their implementation. We also give some intuition on how to choose the kernel of hybrid transforms. Finally, we give heuristics on the topological and geometric information captured by Euler curves and their transforms through their study on synthetic data sets.

V.3.1 Algorithm

In all our experiments, and hence in our implementation, we restrict ourselves to one-critical filtrations. In that case, formulae (V.4) and (V.6) can readily be turned into algorithms computing Euler characteristic profiles and their hybrid transforms. Each algorithm takes as input a grid of size $d_1 \times \ldots \times d_m$ on which the Euler characteristic profile or the hybrid transform

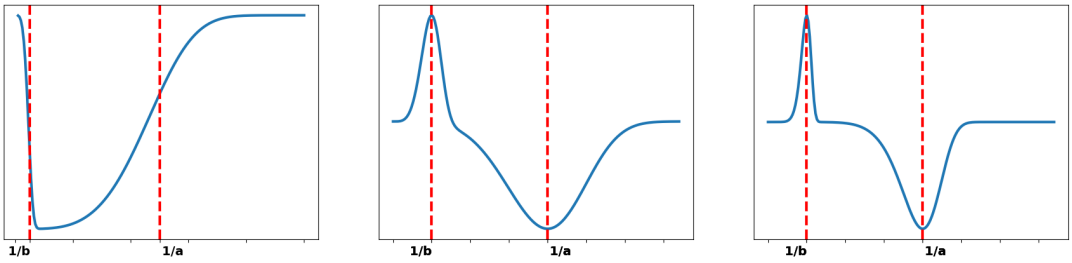
is evaluated. For the Radon transform, we use the fact that $\mathcal{R}_{\mathcal{F}}(\xi, s) = \mathcal{R}_{\mathcal{F}}(\xi/s, 1)$. Our algorithm then takes as input a grid of size $d_1 \times \ldots \times d_m$ on which the map $\eta \in \mathbb{R}_+^{m*} \mapsto \mathcal{R}_{\mathcal{F}}(\eta, 1)$ is evaluated. In any case, the output array of size $d_1 \times \ldots \times d_m$ is an exact sampling of the descriptor. Therefore, our topological descriptors vectorize m-parameter filtrations into $d_1 \times \ldots \times d_m$ arrays that can be used as input to any classical machine learning algorithm.

Complexity. The algorithm computing Euler characteristic profiles with resolution $d_1 \times \ldots \times d_m$ has time complexity $\mathcal{O}(|K| + d_1 \cdot \ldots \cdot d_m)$ in the worst case. The algorithm computing Radon and hybrid transforms with the same resolution has a worst-case time complexity of $\mathcal{O}(|K| \cdot d_1 \cdot \ldots \cdot d_m)$. In comparison, computing a persistence diagram has time complexity $\mathcal{O}(|K|^\omega)$ in the worst case where $2 \leq \omega < 2.373$ is the exponent for matrix multiplication; see [MMS11].

Implementation. A `Python` implementation of our algorithms is freely available online on our GitHub repository: https://github.com/vadimlebovici/eulearning. In practice, our implementation allows for several ways of choosing a grid of sampling. The first method is to provide bounds $[(a_1, b_1), \ldots, (a_m, b_m)]$ and a resolution $d_1 \times \ldots \times d_m$. We then compute a sampling of our descriptors on a uniform discretization of the subset $[a_1, b_1] \times \ldots \times [a_m, b_m] \subseteq \mathbb{R}^m$. This method has the disadvantage of requiring prior knowledge about the data.

For Euler characteristic profiles, the second way is to provide a list $[(p_1, q_1), \ldots, (p_m, q_m)]$ of real numbers $0 \leq p_i < q_i \leq 1$. The algorithm then computes the p_i-th and the q_i-th percentiles of the i-th filtration for each $i = 1, \ldots, m$. Finally, the Euler profiles are uniformly sampled on a $d_1 \times \ldots \times d_m$ grid ranging from the lowest to the highest percentile on each axis. For the Radon and hybrid transforms, the second way consists in providing a list $[p_1, \ldots, p_m]$ of real numbers $0 \leq p_i \leq 1$ and a positive real number α. The algorithm then computes the p_i-th percentiles v_i of the i-th filtration for each $i = 1, \ldots, m$. The integral transforms are uniformly sampled on a $d_1 \times \ldots \times d_m$ grid ranging from 0 to α/v_i on each axis. Note that filtrations have to be positive, which is always satisfied up to translation. This method does not require any prior knowledge of the data but depends on a choice of parameters. More importantly, doing as such is justified for primitive kernels of type $\overline{\kappa} : s \mapsto \exp(-x^p)$ and $\overline{\kappa} : s \mapsto x^p \exp(-x^p)$ in the paragraph below.

Kernel choice. To interpret integral transforms of Euler curves, we set $m = 1$ and compute them on the rectangular function $\chi_{\mathcal{F}} = \mathbf{1}_{[a,b)}$ associated with a persistence diagram with a single point (a, b) with $a < b \in (0, +\infty)$. Recall that the hybrid transform has the simple expression (V.3). Figure 38 shows the hybrid transforms for several kernels. For every $p > 0$, the hybrid transform with primitive kernel $\overline{\kappa} : s \mapsto -\exp(-s^p)$ has a minimum in $\sqrt[p]{\frac{p(\log(b) - \log(a))}{b^p - a^p}}$, which tends to $1/b$ as $p \to \infty$. As a consequence, transforms of this type yield *smoothed* versions of the curve $t \mapsto \chi_{\mathcal{F}}(1/t)$, that is, of an Euler curve with *inverted scales*. Similarly, the hybrid transform with primitive kernel $\overline{\kappa} : s \mapsto -s^p \exp(-s^p)$ has a minimum that tends to $1/a$ and a maximum that tends to $1/b$ as $p \to \infty$, with a spikier aspect as $p \to \infty$. Transforms of this type record the *variations* of the Euler characteristic curve with inverted scales. We refer to the following section for more involved experiments on synthetic data.



(a) $\overline{\kappa}(s) = \exp(-s^4)$ (b) $\overline{\kappa}(s) = s^4 \cdot \exp(-s^4)$ (c) $\overline{\kappa}(s) = s^8 \cdot \exp(-s^8)$

Figure 38: Hybrid transforms of $\chi_{\mathcal{F}} = \mathbf{1}_{[a,b)}$ for several choices of kernel κ

V.3.2 Heuristics for the Euler curves and their transforms

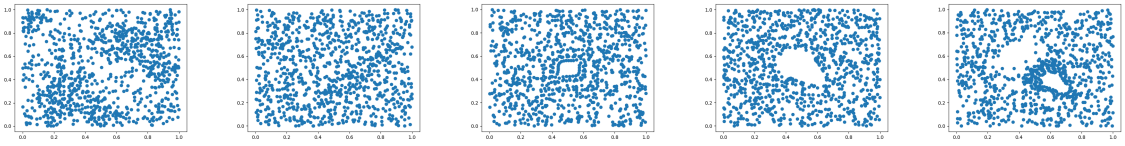
In this section, we assume that $m = 1$ and study the Euler characteristic curves associated with the filtered Čech complex of a point cloud and the hybrid transforms of these curves. We overview how these descriptors can extract information about the topology, geometry, and sampling density of the input data. As already mentioned in Example V.2, we rather use alpha filtration in numerical experiments for computational reasons.

***Topology, sampling*: `ORBIT5K` data set.** While apparently coarse descriptors, Euler characteristic curves allow us to extract relevant scales at which topological differences between two different processes are revealed. We illustrate this fact on the `ORBIT5K` data set from the previous section. We hereby recall its definition.

This data set consists of subsets of a thousand points in the unit cube $[0,1]^2$ generated by a dynamical system that depends on a parameter $\rho > 0$. To generate a point cloud, an initial point (x_0, y_0) is drawn uniformly at random in $[0,1]^2$ and then the sequence of points (x_n, y_n) for $n = 0, \ldots, 999$ is recursively generated via the dynamic:

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

In Figure 39, we illustrate typical orbits for $\rho \in \{2.5, 3.5, 4.0, 4.1, 4.3\}$.



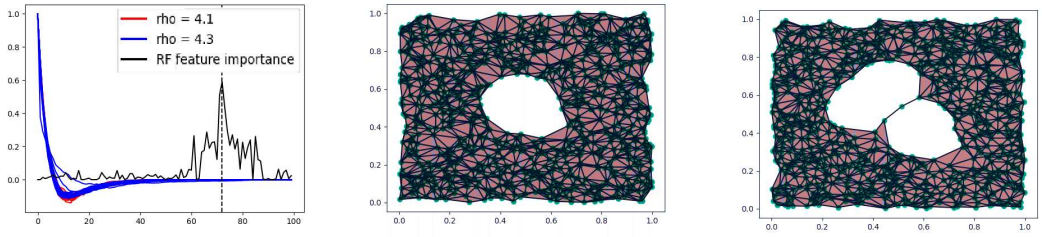
(a) $\rho = 2.5$ (b) $\rho = 3.5$ (c) $\rho = 4.0$ (d) $\rho = 4.1$ (e) $\rho = 4.3$

Figure 39: Examples of point clouds from the `ORBIT5K` data set.

In Figure 40a, we display the Euler characteristic curves for several realizations with parameters $\rho = 4.1$ and $\rho = 4.3$. We also plot the *feature importance* function of a random forest classifier trained on Euler characteristic curves of a small sample of 50 point clouds. In Figures 40b and 40c, we display the alpha complexes for two typical processes truncated at the filtration value corresponding to the largest feature importance. For a large range of

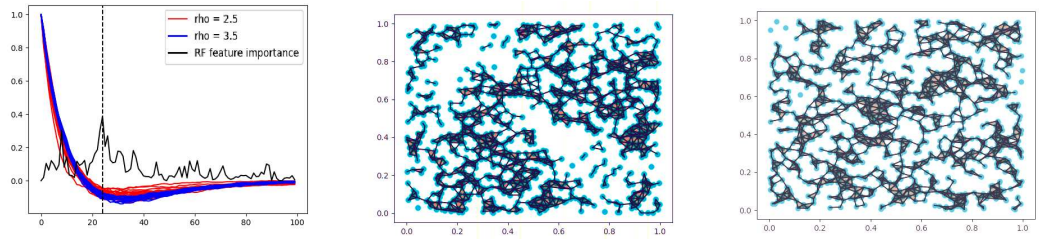
high filtration values—approximately between 60 and 90—, the Euler characteristic curve of each class of process typically differs by one unit since the class with parameter $\rho = 4.3$ has an extra hole. This phenomenon is easily captured by the random forest classifier.

We apply the same methodology to discriminate between $\rho = 2.5$ and $\rho = 3.5$ in Figure 41. The difference between these two classes is more related to the distribution of points than to a persistent topological feature of the point clouds. At the scale selected by the feature importance of the random forest, the alpha complex for $\rho = 2.5$ in Figure 41b tends to have many tiny connected components, while the one for $\rho = 3.5$ is almost connected. We were then able to select a relevant scale at which the difference in the distribution of points is revealed by the topology of the alpha filtration.



(a) ECC and feature importance (b) $\rho = 4.1$ (c) $\rho = 4.3$

Figure 40: `ORBIT5K` classification problem: $\rho = 4.1$ VS $\rho = 4.3$.

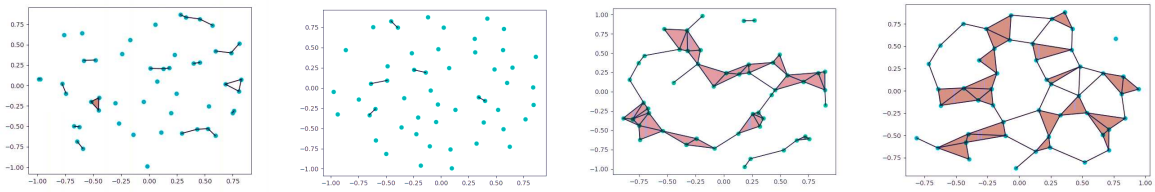


(a) ECC and feature importance (b) $\rho = 2.5$ (c) $\rho = 3.5$

Figure 41: `ORBIT5K` classification problem: $\rho = 2.5$ VS $\rho = 3.5$.

***Sampling*: Poisson and Ginibre point processes.** We perform a similar analysis to discriminate between two types of point processes: a Poisson point process (PPP) and a Ginibre point process (GPP). This setup has been introduced in [OHK18]. The specificity of Ginibre processes lies in repulsive interactions between points. While a standard PPP could have some very small and very large cycles, we expect the GPP to have more medium-sized cycles since points tend to be well dispersed. Ginibre point processes are generated using [DM21]. We classify this toy data set with a random forest classifier and select the two scales corresponding to the most important features of the classifier. In Figure 42, we plot two examples of point clouds together with their alpha complexes at these scales.

We plot Euler curves in Figure 43a. The Euler curves suggest that these classes differ at different scales, as it was visible in Figure 42:

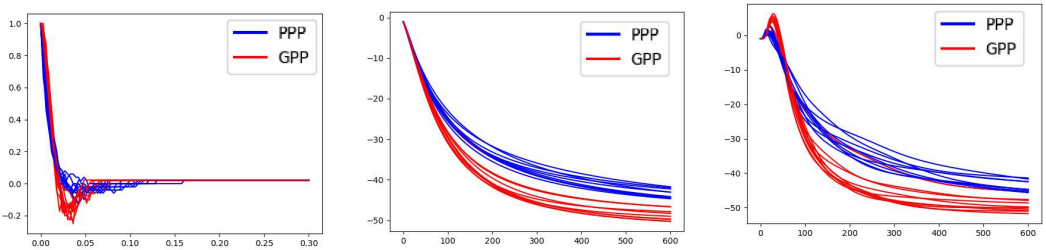


(a) PPP (b) GPP (c) PPP (d) GPP

Figure 42: Examples of alpha complexes on PPP and GPP point clouds at two scales t_1 (Figures (a) and (b)) and t_2 (Figures (c) and (d)) with $t_1 < t_2$.

- The Euler curves of the PPP class decrease in a steeper way. Indeed, a GPP has repulsive interactions between the points. Therefore, the pairwise distance between points tends to be larger and connected components do not die too early.
- The global minimum for the GPP class is lower.
- Compared to curves of the GPP class, the curves of the PPP class tend to stay negative for a longer time. Indeed, PPP allow for very large cycles to exist since there will typically be some large zones without any point, which is proscribed by GPP.

We plot the transforms of these curves for several kernels in Figures 43b and 43c. Choosing the primitive kernel $\overline{\kappa} : s \mapsto \exp(-s)$ emphasises the small scales of the Euler curves in the larger scales of the transform. Such a descriptor separates well the two classes due to the earlier death of connected components for the PPP class. The primitive kernel $\overline{\kappa} : s \mapsto \exp(-s^4)$ also extracts this information. In addition, it has a higher global maximum for the GPP class that also enables distinction between the two classes. This maximum is created by the global minimum of the Euler curves. This experiment is a piece of evidence that this kernel carries more information than the exponential kernel and will therefore be preferred for applications.



(a) ECC (b) HT, $\overline{\kappa}(s) = \exp(-s)$ (c) HT, $\overline{\kappa}(s) = \exp(-s^4)$

Figure 43: Euler characteristic curves and their transforms for PPP VS GPP data set

***Geometric features, sampling*: different samplings on a manifold.** We now show an experiment where we can illustrate how our various descriptors can discriminate between samplings and characterize the shape of a manifold. We consider two set-ups. The first set-up consists of clouds of 500 points sampled in two different ways on a torus embedded in $\mathbb{R}^3$. The first sampling is a uniform sampling [DHS+13]. The second is a non-uniform sampling where we draw (θ, φ) uniformly in $[0, 2\pi]^2$ and obtain a point on the torus through the embedding $\Psi_{\mathbb{T}^2} : (\theta, \varphi) \mapsto (x_1, x_2, x_3)$ where:

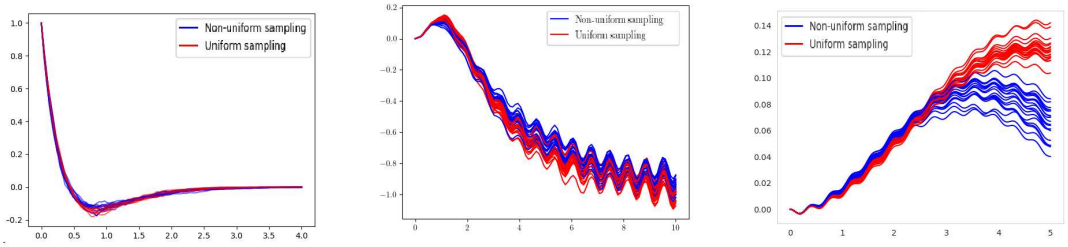
$$\begin{cases} x_1 = (2 + \cos(\theta))\cos(\varphi), \\ x_2 = (2 + \cos(\theta))\sin(\varphi), \\ x_3 = \sin(\theta). \end{cases}$$

The second set-up consists of clouds of 500 points drawn in two ways on the unit sphere of $\mathbb{R}^3$. The first sampling is uniform. The second sampling is a non-uniform sampling where we draw θ uniformly in $[0, \pi]$ and φ according to a normal distribution centred on π. We obtain a point on the sphere via the classical spherical coordinates parametrization $\Psi_{\mathbb{S}^2} : (\theta, \varphi) \mapsto (x_1, x_2, x_3)$ where:

$$\begin{cases} x_1 = \sin(\theta)\cos(\varphi), \\ x_2 = \sin(\theta)\sin(\varphi), \\ x_3 = \cos(\theta). \end{cases}$$

In Figures 44a and 44b, we show the Euler curves and their hybrid transforms with primitive kernel $\overline{\kappa} : s \mapsto \cos(s)$ for these two classes of samplings on the torus. Up to a rescaling, this corresponds to a Fourier sine transform. In Figure 44c, we show the hybrid transforms for our two classes of samplings on the sphere.

In both cases, Euler curves associated with data drawn on the same manifold all have the same profile, with a minimum value that tends to be lower for the uniform sampling. Similarly, the oscillations of the transforms are in phase and have the same amplitude. However, from one manifold to another, the phase and amplitude of the oscillations of the transforms differ significantly. This suggests that they are related to global quantities and are signatures of the support manifold. In contrast, the samplings show up in the vertical shifts of the oscillations of the transforms. This interpretation allows us to go beyond the classical signal/noise dichotomy developed in Section III in the persistence diagrams. Although it makes no doubt that this sampling information can be retrieved from low-persistence features, it is still unclear how to read it from a persistence diagram. We claim this is another step towards a more thorough analysis of the geometric quantities involved in the low-persistence features.

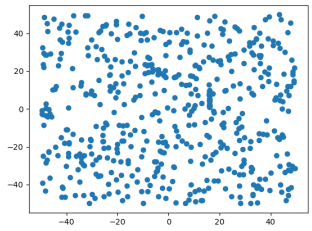


(a) ECC, torus data (b) HT, torus data (c) HT, sphere data

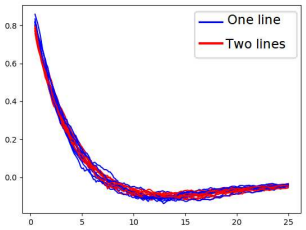
Figure 44: ECC and HT, two different samplings on a torus and a sphere

***Geometric features, sampling*: two different patterns in noise.** In this final illustrative experiment, we try to distinguish patterns in a heavy clutter noise. One class has one line hidden in the noise, while the other has two. Each line will induce a very dense zone creating early dying connected components. In Figure 45, we plot two examples of point clouds, the Euler curves of each class, and their hybrid transform with primitive kernel $\overline{\kappa} : s \mapsto \exp(-s^4)$. We also provide PCA plots of these two descriptors. The difference between the two classes is

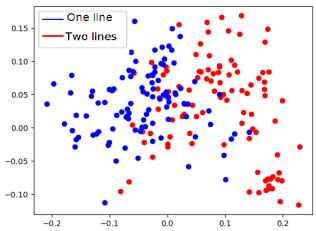
visible at the beginning of the Euler characteristic curves. However, looking at the full curve does not allow us to correctly see this difference, as shown by the PCA plot. On the contrary, the transform puts a strong emphasis on the beginning of the Euler curves, leading to a direct linear separation of the two classes. As a final sanity check, we ran a k-means algorithm to cluster between the two classes and reached an accuracy of 99% for the hybrid transforms and only 52.5% for the Euler curves.



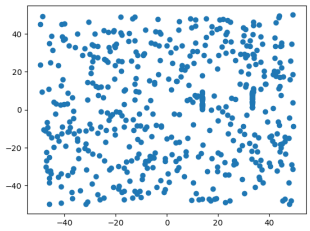
(a) raw data, one hidden line



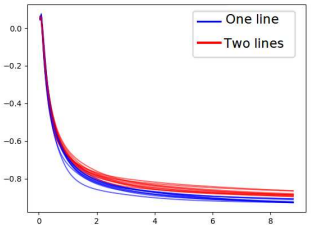
(b) ECC



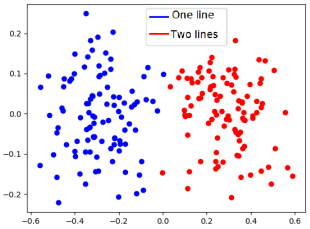
(c) PCA on ECC



(d) raw data, two hidden lines



(e) HT, $\overline{\kappa}(s) = \exp(-s^4)$



(f) PCA on HT

Figure 45: Pattern hidden in clutter noise

V.4 Experiments

In this section, we present all quantitative experiments conducted on synthetic and real-world point cloud data and on real graph data sets. Material to reproduce our experiments is available online on our GitHub repository: https://github.com/vadimlebovici/eulearning.

V.4.1 Curvature regression

We consider a set-up from [BHPW20] where we draw 1000 points uniformly at random on the unit disk of a surface of constant curvature K and try to predict K in a supervised fashion. Recall that if $K > 0$ (resp. $K = 0$, $K > 0$), the corresponding surface is a sphere (resp. the Euclidean plane, the hyperbolic plane). We observe 101 samples from space with curvature $[-2, -1.96, \ldots, 1.96, 2]$ and validate our model on a testing set of 100 point clouds sampled from space with random curvature drawn uniformly in $[-2, 2]$. We compare the R^2 score in Table 13 with that of the original paper, which uses persistent landscapes (PL) along with a support vector regressor (SVR) and with Persformer [RCB21]. Note that since we are trying to tackle a regression problem, we use an SVR or a random forest regressor to predict the curvature from our vectorization.

First, we remark that the ECC descriptor combined with a random forest has an accuracy comparable to the state-of-the-art. We also remark that taking a transform does not improve the regression accuracy when considering a robust classifier such as RF but does improve

Method	PL+SVR	Persformer	ECC+SVR	ECC+RF	HT+SVR	HT+RF
R^2 score	0.78	**0.94**	0.70	0.93	0.79	0.89

Table 13: R^2 score for curvature regression data

the accuracy when using a linear regressor (SVR). Note that hybrid transforms combined with a linear regressor have an accuracy similar to that of persistent landscapes. However, persistent landscapes require the computation of the entire persistence diagrams, while hybrid transforms bypass this costly operation.

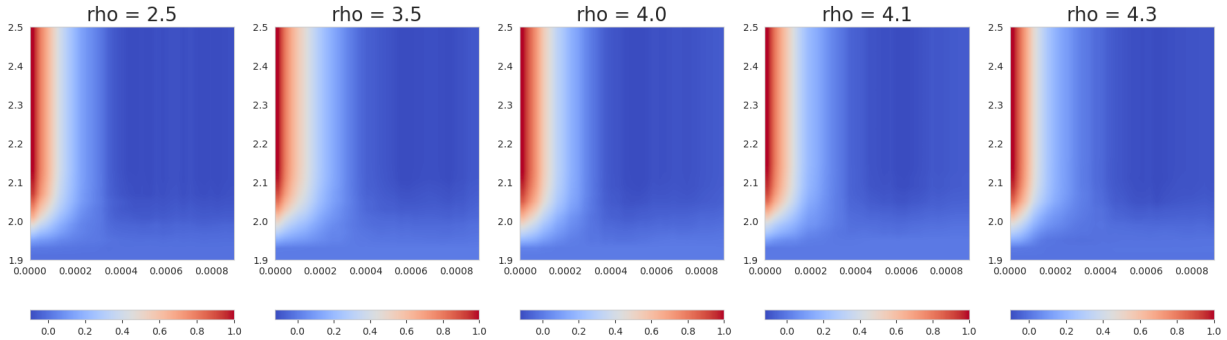
V.4.2 `ORBIT5K` data set

Supervised setting. Here, we perform a supervised analysis of the `ORBIT5K` data set introduced in Section V.3.2. Given an orbit, we try to predict the value of the parameter ρ, which takes value in $\{2.5, 3.5, 4.0, 4.1, 4.3\}$. We generate 700 training and 300 testing orbits for each class. We compare our score with standard classification methods in Table 14. The results are averaged over ten runs. PWG-K, SW-K and PF-K are kernel methods on persistence diagrams taken respectively from [KHF16, CCO17, LY18]. Perslay and Persformer are two methods that use a neural network architecture to vectorize persistence diagrams [CCI+20, RCB21]. The Euler characteristic curves and one-parameter hybrid transforms (HT1) are computed on the alpha filtration of the point cloud. The Euler characteristic surfaces, the two-parameter Radon transform (RT) and hybrid transforms (HT2) are computed using a function-alpha filtration associated with a kernel density estimator post-composed with a decreasing function. The decreasing function is $x \mapsto -x$ for the ECSs and $x \mapsto \exp(-x^2)$ for the HTs. All descriptors have a resolution of 900 (hence of 30×30 for two-parameter ones) and were trained with an XGBoost classifier [CG16]. We select the hyperparameters of our descriptors by cross-validation:

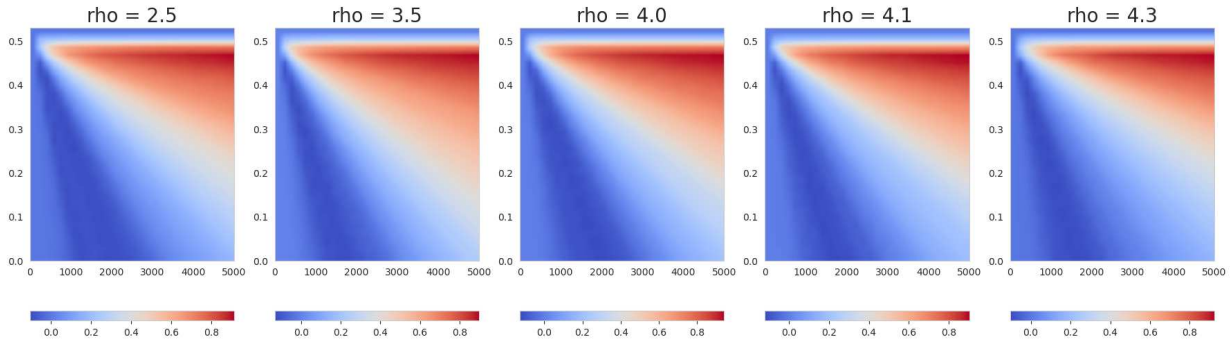
- For the ECC, the quantiles (see *Implementation* in Section V.3.1) are selected in $\{(0.1, 0.9), (0.2, 0.8), (0.3, 0.7)\}$.
- For the ECS, the quantiles are selected in the same set as for the ECC for both parameters.
- For the HT1, the range is selected in $\{[0, 50], [0, 100], [0, 500], [0, 1000]\}$ and the primitive kernel $\overline{\kappa}$ in $\{s \mapsto \exp(-s^4),\ s \mapsto s^4 \exp(-s^4), s \mapsto s^8 \exp(-s^8)\}$.
- For the HT2, the primitive kernel and the range for the first parameter are the same as for the HT1, and the range for the second parameter is selected in $\{[0, 50], [0, 80], [0, 100], [0, 500]\}$.
- For the RT, the ranges are selected in the same set as the HT2 for both parameters.

We show in Figure 46 some examples of each descriptor renormalized by the number of points for the classes $\rho = 2.5$ and $\rho = 4.3$, where the HT2 is computed with $\overline{\kappa} : s \mapsto s^4 \exp(-s^4)$.

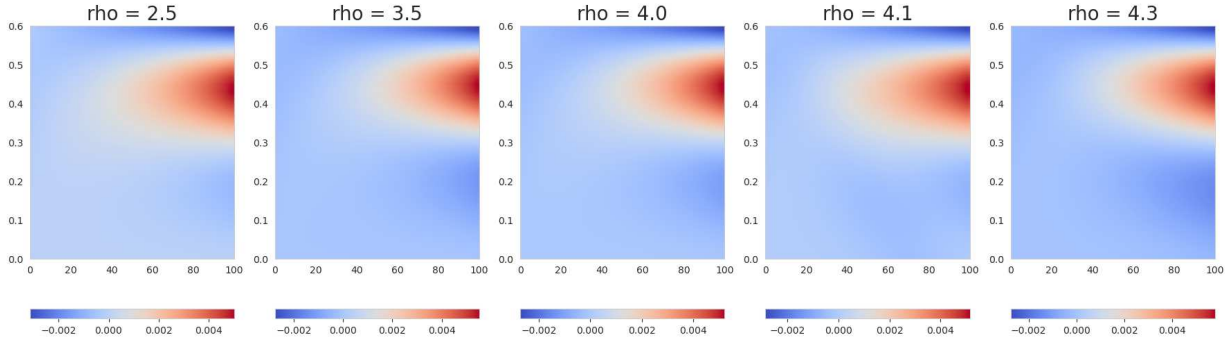
One-parameter descriptors have accuracy similar to kernel methods on persistence diagrams at a reduced computational cost, while two-parameter descriptors compete with neural network-based vectorization methods. We make our claims on computational times more precise in Section V.4.5.



(a) ECS



(b) RT



(c) HT2

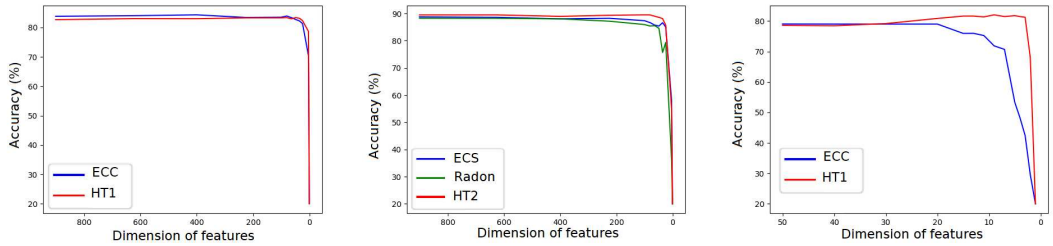
Figure 46: Examples of 2D descriptors

Method	PWG-K	SW-K	PF-K	Perslay	Persformer
Accuracy	76.6 ± 0.7	83.6 ± 0.9	85.9 ± 0.8	87.7 ± 1.0	91.2 ± 0.8
Method	ECC + XGB	HT1 + XGB	**ECS + XGB**	RT + XGB	HT2 + XGB
Accuracy	83.8 ± 0.5	82.8 ± 1.4	**91.8 ± 0.4**	90.5 ± 0.4	89.9 ± 0.5

Table 14: Classification scores for the `ORBIT5K` data set

Ablation study. We also study the role of the dimension of the feature vector in the supervised classification task. The results are shown in Figure 47. When plugging a random forest classifier, all descriptors are robust to a decrease in the size of the feature vector. However, hybrid transforms seem to maintain a competitive accuracy for low-dimensional

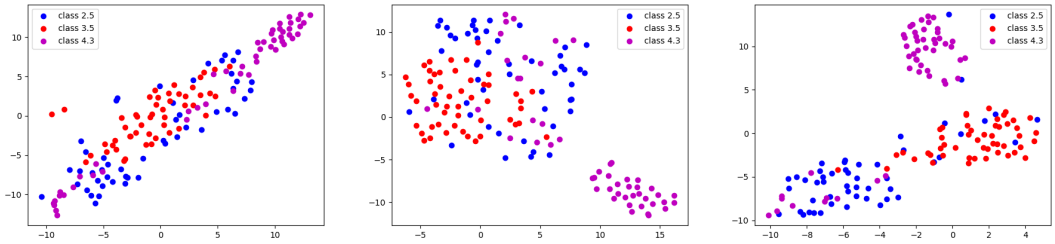
features, especially the two-parameter ones. When using an SVM classifier for the one-parameter descriptors, the gain from considering a hybrid transform is clear, and the accuracy of the SVM benefits from this strong dimension reduction. Evaluating hybrid transforms at only three values of $\xi \in \mathbb{R}_+^*$ yields feature vectors achieving approximately 80% accuracy, demonstrating the compression properties of this tool.



(a) one-parameter, RF classifier (b) two-parameter, RF classifier (c) one-parameter, SVM classifier

Figure 47: Accuracy with respect to feature dimension.

Unsupervised setting. We consider several unsupervised classification tasks on the same data set. We consider 50 point clouds for each choice of $\rho \in \{2.5, 3.5, 4.3\}$. We map all descriptors in $\mathbb{R}^2$ using a tSNE dimension reduction [VdMH08] and report the results in Figure 48. Here, hybrid transforms differ from the other two methods and succeed in adequately separating the three classes.



(a) ECS (b) RT (c) HT2

Figure 48: tSNE of our descriptors computed on several classes of the `ORBIT5K` data set.

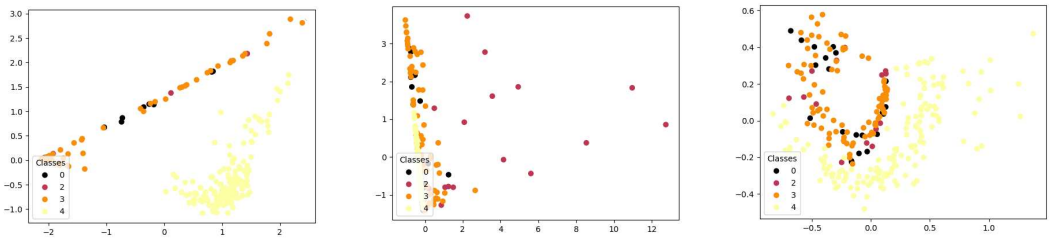
V.4.3 Sidney object recognition data set

The Sidney urban objects recognition data set consists of 3D point clouds of everyday urban road objects scanned with a LIDAR [DDQHD13] traditionally used for multi-class classification. Likewise to Section V.4.2, all descriptors are computed using a function-alpha filtration associated with a kernel density estimator post-composed with a decreasing function.

Unsupervised setting. In Figure 49, we show a PCA of the ECSs and HTs on the classes *4-wheeler vehicles* (labelled 0), *buses* (2), *cars* (3), and *pedestrians* (4). In this case, the ECSs separate the class of pedestrians from all the vehicle classes. The same separation is achieved

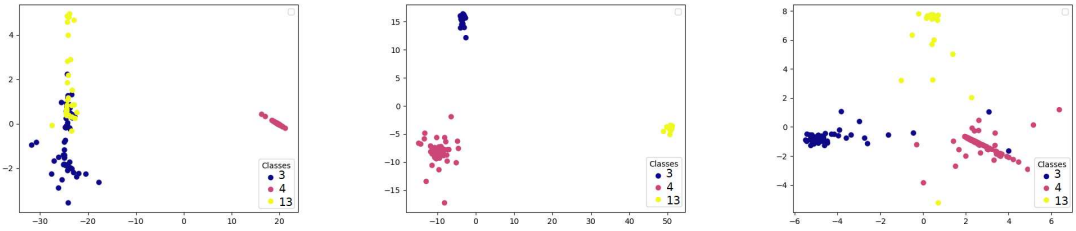
by the HTs with primitive kernel $\overline{\kappa} : s \mapsto s^4 \exp(-s^4)$. In contrast, HTs with primitive kernel $\overline{\kappa} : s \mapsto \exp(-s^4)$ separate buses from other classes. These experiments illustrate the flexibility provided by a broad choice of kernels for the hybrid transforms.

Supervised setting. Even more striking are the experiments from Figure 50. We perform a Linear Discriminant Analysis for classes *cars* (3), *pedestrians* (4), and *vans* (13) to embed the HTs and ECSs in $\mathbb{R}^2$. All the classes are separated for the RTs and the HTs with primitive kernel $\overline{\kappa} : s \mapsto s^4 \exp(-s^4)$. In comparison, the ECSs only manage to separate the pedestrian class from the two motor-vehicle classes.



(a) ECS (b) HTs, $\overline{\kappa}(s) = \exp(-s^4)$ (c) HTs, $\overline{\kappa}(s) = s^4 \exp(-s^4)$

Figure 49: PCA plots of ECSs and HTs for the Sidney object recognition data set.



(a) ECS (b) HTs, $\overline{\kappa}(s) = s^4 \exp(-s^4)$ (c) RTs

Figure 50: LDA plots of ECSs, HTs, and RTs for the Sidney object recognition data set.

V.4.4 Graph data

We have applied our method to the supervised classification of graph data. To build sublevel-sets filtrations of graphs, we consider the heat-kernel signature from the previous section. We set the parameters to $t = 1$ and $t = 10$. In addition, we consider the $1/2-$Ricci and Forman curvatures [SSG+18], centrality, and edge betweenness on connected graphs. In addition, some data sets (PROTEINS, COX2, DHFR) come with functions defined on the graph nodes. We can use several combinations of these functions to define sublevel-sets filtrations of graphs and compute Euler characteristic profiles (ECP), Radon transforms (RT) and hybrid transforms (HTn).

For this set of experiments, we cross-validate over several combinations of the filtration functions proposed above, several truncations of the vectorization (which had little impact in practice), and a primitive kernel chosen among $\{s \mapsto \cos(s), s \mapsto \cos(s^2), s \mapsto \exp(-s^4), s \mapsto$

Method	MUTAG	COX2	DHFR	PROTEINS	COLLAB	IMDB-B	IMDB-M	NCI1
SV	88.2(0.1)	78.4(0.4)	78.8(0.7)	72.6(0.4)	79.6(0.3)	74.2(0.9)	49.9(0.3)	71.3(0.4)
RetGK	90.3(1.1)	**81.4(0.6)**	81.5(0.9)	**78.0(0.3)**	81.0(0.3)	71.9(1.0)	47.7(0.3)	**84.5(0.2)**
FGSD	**92.1**	-	-	73.4	80.0	73.6	**52.4**	79.8
GIN	90(8.8)	-	-	76.2(2.6)	80.6(1.9)	**75.1(5.1)**	52.3(2.8)	82.7(1.6)
Perslay	89.8(0.9)	80.9(1.0)	80.3(0.8)	74.8(0.3)	76.4(0.4)	71.2(0.7)	48.8(0.6)	73.5(0.3)
Atol	88.3(0.8)	79.4(0.7)	82.7(0.7)	71.4(0.6)	**88.3(0.2)**	74.8(0.3)	47.8(0.7)	78.5(0.3)
ECC 1D	87.2(0.7)	78.1(0.2)	79.4(0.5)	74.7(0.4)	77.3(0.2)	72.4(0.4)	48.5(0.3)	74.4(0.2)
HT 1D	87.4(0.8)	78.1(0.2)	77.9(0.4)	73.3(0.4)	78.2(0.2)	73.9(0.4)	49.7(0.4)	73.9(0.2)
ECV	90.0(0.8)	80.3(0.4)	82.0(0.4)	75.0(0.3)	78.3(0.1)	73.3(0.4)	48.7(0.4)	76.3(0.1)
RT	87.3(0.6)	79.7(0.4)	81.3(0.4)	75.4(0.4)	77.5(0.2)	74.0(0.5)	50.2(0.4)	75.6(0.2)
HT nD	89.4(0.7)	80.6(0.4)	**83.1(0.5)**	75.4(0.4)	77.6(0.2)	74.7(0.5)	49.9(0.4)	76.4 (0.2)

Table 15: Mean accuracy and standard deviation for graph data.

$s^4 \exp(-s^4)\}$ for HTn. We report our scores in Table 15. The first four methods are state-of-the-art classification methods on graphs that use kernels or neural networks. We report the scores from the original papers, [TVH19, ZWX+18, VZ17, XHLJ19]. Perslay [CCI+20], and Atol [RCL+21] are topological methods that transform the graphs into persistence diagrams using HKS functions. It is known that Atol performs especially well on large data sets (both in terms of number of data and graphs size), i.e., COLLAB and NCI1. Still, we reach a similar to better accuracy for all the other data sets.

Besides highly competitive classification scores, our method has two advantages over the other topological methods. First, we bypass the computation of persistence diagrams and thus classify with lower computational cost; see Sections V.3.1 and V.4.5. Second, as opposed to other invariants such as multi-parameter persistent images [CB20], our method naturally generalizes to m-parameter persistence with $m \geq 3$ at a very low computational cost. To our knowledge, this is the first time a topology-based method uses more than 3 filtration parameters. This results in an increase in accuracy since each filtration function leverages information on the graph-data structures.

Note that the methods SV, FGSD, and GIN do not average ten times and rather consider a single 10-fold sample which can slightly boost their accuracies.

V.4.5 Timing

We choose to compare the computational cost of our methods to that of persistence images as they appear to be a faster vectorization method than persistence kernels and persistence landscapes, especially in a multi-persistence setting; see [CB20, Table 2].

Constant resolution. We report in Table 16 the time to compute our descriptors and persistent images on the full `ORBIT5K` data set with a fixed resolution of 900. Likewise to the `Gudhi` library, our method uses the *simplex tree* data structure to represent simplicial complexes, see [BM14]. We report computation times for precomputed simplex trees [7] using the `Gudhi` library, [Rou15]. All descriptors are computed using the parameters achieving the highest accuracy for the classification task; see Section V.4.2. Persistence images are computed with the `Gudhi` library for one-parameter filtrations and with the `MMA` package for

[7] Note that computing simplex trees takes around 66s in the one-parameter setting and around 420s in the two-parameter setting; the difference lies in the cost of computing a codensity estimator on point clouds.

two-parameter filtrations [LCS22] with default parameters and the same resolution as our two-parameter descriptors, i.e., 30×30. To compute persistence images, one first needs to compute the persistence diagrams in the one-parameter case or persistence modules approximations in the two-parameter case [LCB22, Section 3]. We include these additional costs in the computational times of persistent images. However, the time to compute the PI1 descriptor on the full `ORBIT5K` data set breaks down to 5 seconds to compute the persistence diagrams and 134 seconds for the persistence images themselves.

ECC	HT1	PI1	ECS	Radon	HT2	PI2
16	719	139	144	119	805	2034

Table 16: Computation times (s) for `ORBIT5K` with constant resolution.

As expected from the time complexities of the algorithms (Section V.3.1), Euler characteristic profiles and Radon transforms are at least ten times faster than persistence images, and hybrid transforms are four times faster in the two-parameter case. One-parameter hybrid transforms may appear costly to compute, but this will be mitigated in the next paragraph. Finally, we point out that we implemented our tools in `Python` and not in `C++`, which is very likely to result in longer computation times. On the contrary, persistence images in one and two parameters both benefit from a `C++` implementation.

Constant accuracy. We report in Table 17 the time to compute our descriptors on the full `ORBIT5K` data set with the lowest resolution before accuracy drop-out as reported in Figure 47. More precisely, we chose the lowest possible resolutions to ensure a classification accuracy of 82% for one-parameter descriptors and of 89% for two-parameter descriptors, that is, a resolution of 30 for ECC, of 9 for HT1, of 20×20 for ECS and Radon, and of 6×6 for HT2. Other parameters remain unchanged. The interest in using hybrid transforms over Euler characteristic profiles is now clear: the concentration of information provided by hybrid transforms makes it possible to classify the data set with feature vectors of reduced dimension, which considerably speeds up the computations.

ECC	HT1	ECS	Radon	HT2
16	5	135	45	69

Table 17: Computation times (s) for `ORBIT5K` with smallest resolution before accuracy drop-out.

V.4.6 Take-home message

The experiments from this section suggest that Euler characteristic profiles are very powerful descriptors since they allow for state-of-the-art accuracy when coupled with a robust classifier (XGB or RF) at a very competitive computational cost. On the one hand, Radon transforms show accuracy and computational complexity very similar to Euler characteristic profiles. On the other hand, hybrid transforms have similar accuracy but are more costly to compute, especially in the one-parameter setting; see Table 16. The motivation to use hybrid transforms is two-fold:

- In an unsupervised setting or when plugging a linear classifier, the lack of diversity in Euler characteristic profiles and Radon transforms can be detrimental to the separation of classes. In contrast, hybrid transforms are competitive descriptors in such tasks due to the wide diversity in the choice of kernels and their sensitivity to slight variations in Euler characteristic profiles.
- Hybrid transforms provide a very powerful compression of the signal from the Euler profile (Figure 47) at a very low computational cost (Table 17). This makes hybrid transforms a very robust descriptor combining dimension reduction and feature extraction.

On the theoretical side, multi-parameter hybrid transforms benefit from their expression as one-parameter ones (Lemma V.9). This allows us to prove almost sure convergence results under some mild assumptions in Section V.6.

V.4.7 Extensions

We have validated our method on simplicial complexes built on point clouds and graph data. Nonetheless, the methodology described in this section can be extended into two directions.

First, when dealing with images or 3D volumes, it is common to build cubical complexes from data. In this context, Euler characteristic curves have been used as a vectorization of the data in [SZ21, JKN20]. As there are a vast number of filtration functions one can consider on images, it is worth investigating the predictive power of the Euler characteristic profiles in this setting. While several applications are considered in [BSA+22, DG22], a thorough benchmark study against other persistence methods and state-of-the-art image processing methods is still missing. Moreover, Radon transforms and hybrid transforms have still not been studied in this context.

Second, the methodology developed here applies to filtrations $\mathcal{F} = (\mathcal{F}_t)_{t\in\mathbb{R}^m}$ that are not necessarily non-decreasing with respect to inclusions. This extends the potential range of applications of our tools, notably to the study of time-varying simplicial complexes, as done in [XATZ22].

V.5 Stability properties

The success of topological data analysis inherits from the stability theorem for persistence diagrams from [CSEH07], in particular Theorem I.15. Loosely speaking, it means that under mild assumptions, small changes in the filtration function imply small changes in the persistence diagram. Such results are key to designing consistent estimators in statistical analysis; see, for instance, [BMT17]. More recently, [ST20] have derived a stability result for the p-Wasserstein distance between persistence diagrams, and several stability results for Euler characteristic tools have been derived in [CMT22, DG22].

In this section, we state stability results for our topological descriptors. Our results compare the L^1 norm between Euler characteristic profiles to the signed 1-Wasserstein distance between their signed barcodes. As a corollary, we bound the L^q norms of Radon transforms and hybrid transforms by the same quantity. The notions of signed barcodes and of signed 1-Wasserstein distance are introduced in [OS21] and recalled below. We follow the same conventions as in [OS21, Section 2] for the definitions of multisets and bijections between them. All the results of this section are proven in Section V.7.

Signed 1-Wasserstein distance. The distance we use to state our stability results is defined on the class of *finitely presented* functions over $\mathbb{R}^m$, that is, which can be written as a finite $\mathbb{Z}$-linear combination of indicator functions $\mathbf{1}_{Q_u}$ for some $u \in \mathbb{R}^m$. These functions include Euler characteristic profiles of one-critical filtrations. A *decomposition* of a finitely presented function φ is a couple (η^+, η^-) of finite multisets of points in $\mathbb{R}^m$ such that:

$$\varphi = \sum_{u \in \eta^+} \mathbf{1}_{Q_u} - \sum_{v \in \eta^-} \mathbf{1}_{Q_v}.$$

Such a decomposition always exists, and there is a unique $\overline{\mathcal{B}} = (\eta^+, \eta^-)$ such that $\eta^+ \cap \eta^- = \varnothing$, called the *signed barcode of* φ; see [OS21, Proposition 13].

Let $\mathcal{C}$ and $\mathcal{C}'$ be two finite multisets of points in $\mathbb{R}^m$ with the same cardinality and $h : \mathcal{C} \to \mathcal{C}'$ be a bijection between them. The *cost* of h is the real number $\mathrm{cost}(h) = \sum_{u \in \mathcal{C}} \|u - h(u)\|_1$. For any two finitely presented functions φ and φ' with respective signed barcodes (η^+, η^-) and (η'^+, η'^-), the *signed 1-Wasserstein distance* between them is:

$$\widehat{d}_1(\varphi, \varphi') = \inf \left\{ \varepsilon > 0 : \exists \text{ bijection } h : \eta^+ \cup \eta'^- \to \eta^- \cup \eta'^+ \text{ with } \mathrm{cost}(h) \leq \varepsilon \right\}.$$

Hence, one has $\widehat{d}_1(\varphi, \varphi') \in [0, +\infty]$. Note that bijections do not allow for unmatched bars, as it is common in the persistence literature. In loc. cit., the signed 1-Wasserstein distance is defined on signed barcodes. Our definition is essentially equivalent since signed barcodes are in one-to-one correspondence with finitely presented functions up to forgetting the order in the multisets.

Stability results. We can state our first stability result. The case $m = 1$ is well known for 1-Wasserstein distance on persistence diagrams; see [CMT22, Lemma 4.10], [DG22, Proposition 3.2].

Proposition V.12. *Let $\mathcal{F}$ and $\mathcal{F}'$ be two finitely generated m-parameter filtrations of simplicial complexes $\mathcal{K}$ and $\mathcal{K}'$ respectively. For any $M > 0$, we have that*

$$\|\chi_{\mathcal{F}} - \chi_{\mathcal{F}'}\|_{1,M} \leq (2M)^{m-1} \, \widehat{d}_1(\chi_{\mathcal{F}}, \chi_{\mathcal{F}'}).$$

In particular, if $m = 1$:

$$\|\chi_{\mathcal{F}} - \chi_{\mathcal{F}'}\|_1 \leq \widehat{d}_1(\chi_{\mathcal{F}}, \chi_{\mathcal{F}'}).$$

Combined with the results of the previous paragraph, the above proposition links the L_1 distance between Euler characteristic profiles to a classical distance of persistence theory. Moreover, these stability results carry over to our other descriptors, as stated in the Corollary V.13 below. Let K be a compact subset of $\mathbb{R}_+^{m*}$. For $q \in [1, \infty]$, we consider the norms on functions $\mathcal{R} : \mathbb{R}_+^{m*} \times \mathbb{R} \to \mathbb{R}$ defined by:

$$\|\mathcal{R}\|_{L_K^{q,1}} = \begin{cases} \left(\displaystyle\int_K \left(\int_{\mathbb{R}} |\mathcal{R}(\xi, s)| \, \mathrm{d}s \right)^q \mathrm{d}\xi \right)^{1/q} & \text{for } q \in [1, \infty), \\ \displaystyle\sup_{\xi \in K} \int_{\mathbb{R}} |\mathcal{R}(\xi, s)| \, \mathrm{d}s & \text{for } q = \infty. \end{cases} \tag{V.8}$$

Corollary V.13. *Let K be a compact subset of $\mathbb{R}_+^{m*}$ and $q \in [1,\infty]$. Let $\mathcal{F}$ and $\mathcal{F}'$ be one-critical m-parameter filtrations of simplicial complexes $\mathcal{K}$ and $\mathcal{K}'$ respectively. Let $\kappa \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$. There exists a constant $C_{K,q}$ depending only on K and q such that:*

$$\|\mathcal{R}_{\mathcal{F}} - \mathcal{R}_{\mathcal{F}'}\|_{L_K^{q,1}} \leq C_{K,q}\, \widehat{d}_1(\chi_{\mathcal{F}}, \chi_{\mathcal{F}'})\,,$$

$$\|\psi_{\mathcal{F}}^{\kappa} - \psi_{\mathcal{F}'}^{\kappa}\|_{L_K^{q}} \leq C_{K,q}\, \|\kappa\|_\infty\, \widehat{d}_1(\chi_{\mathcal{F}}, \chi_{\mathcal{F}'})\,.$$

In the case of filtrations of the same simplicial complex, we can turn the above results into stability results with respect to L^1 norms on filtration functions, as stated by the lemma below. It has already been formulated in a slightly different form in [DG22, Proposition 3.4]. Let $\mathcal{K}$ be a finite simplicial complex, and $f : \mathcal{K} \to \mathbb{R}^m$ a non-decreasing map. We define the 1-*norm* of f as $\|f\|_1 = \sum_{\sigma\in\mathcal{K}} \|f(\sigma)\|_1$.

Lemma V.14. *Let $\mathcal{K}$ be a finite simplicial complex and $f, g : \mathcal{K} \to \mathbb{R}^m$ be non-decreasing maps. We have that*

$$\widehat{d}_1(\chi_f, \chi_g) \leq \|f - g\|_1.$$

The above lemma clarifies the robustness of our descriptors with respect to perturbations of filtrations defined on a fixed simplicial complex. This includes, for instance, density estimators on point clouds or Ricci curvature and HKS functions on graphs. The fact that these descriptors are controlled by the L^1 distance and not the L^∞ distance between the functions is an indicator of their sensitivity to the underlying geometry. Persistent images [AEK+17] share this property, while persistence landscapes [B+15, Vip20] do not, as they are controlled by the L^∞ distance between functions.

V.6 Statistical properties

In this section, we provide statistical guarantees for our descriptors computed on a random sample, as the sample size tends to infinity.

V.6.1 Limit theorems for one-parameter hybrid transforms

This section is devoted to limit theorems for the hybrid transforms of the Čech complex of an i.i.d. sample in $\mathbb{R}^d$. Theorem V.15 is a pointwise law of large numbers, while Theorem V.17 states a functional central limit theorem for the hybrid transforms of compactly supported kernels. The purpose of this section is two-fold: we state that under some mild assumptions, hybrid transforms are universal in the sense that they converge to an object that depends only on the kernel, the filtration, and the sampling scheme. In addition, we illustrate how information on the sampling can be extracted from the limiting object in Theorems V.15 and V.17. This shows that if the number of sample points is large enough, hybrid transforms are relevant tools to perform classification tasks on point clouds. These results back up the experiments of the previous sections. They are proven in Section V.7.

Theorem V.15. *Let $X_1, \ldots, X_n$ be an i.i.d. sample drawn according to an a.e. continuous bounded Lipschitz density g on $\mathbb{R}^d$. Consider a sequence $(r_n)_{n\in\mathbb{N}}$ such that $nr_n^d \to 0$ and $n^{k+2} r_n^{d(k+1)} \to \infty$ as $n \to \infty$ for all k in $[\![0, d-1]\!]$. We denote by $\mathcal{F}_n$ the Čech filtration associated with the rescaled sample $\frac{1}{r_n}(X_i)_{i=1}^n$. Let $T, a > 0$ and $\kappa \in L^1(\mathbb{R})$. Further assume*

that κ is supported on $[0,T]$. Then there exist functions $A_0,\dots,A_{d-1}$ on $\mathbb{R}_+^*$ that depend only on $\overline{\kappa}$ such that for every $\xi > a$,

$$\frac{1}{n^{k+2}r_n^{d(k+1)}}\cdot\psi_{\mathcal{F}_n}^{\kappa}(\xi) \underset{n\to\infty}{\longrightarrow} \sum_{k=0}^{d-1}\frac{(-1)^k}{(k+2)!}\cdot A_k(\xi)\cdot\int_{\mathbb{R}^d} g^{k+2}(x)\,\mathrm{d}x \quad a.s..$$

This theorem is a direct consequence of the limit theorems derived in [Owa22]. It is a key assumption that we are in the so-called *sparse regime*, that is, $nr_n^d \to 0$. In order to make this law of large numbers more comprehensible, we make a further assumption that we are in the so-called *divergence regime*, that is $n^{k+2}r_n^{d(k+1)} \to \infty$ for all $k \in [\![0, d-1]\!]$. The sequence defined by $r_n = n^{-\alpha}$ for $\frac{1}{d} < \alpha < \frac{1}{d} + \frac{1}{d^2}$ verifies these two assumptions. Similar results can be derived for other subcases of the sparse regime, i.e., the Poisson regime $n^{k+2}r_n^{d(k+1)} \to c > 0$ and the vanishing regime $n^{k+2}r_n^{d(k+1)} \to 0$.

This theorem shows that the pointwise limit of the hybrid transform depends on the sampling only through the quantities $\int_{\mathbb{R}^d} g^{k+2}$ for $k = 0, 1, \dots, d-1$ and they can therefore discriminate between different samplings as soon as n is large enough. In addition to this law of large numbers, a finer limit result for the Euler characteristic curve is proven in [KRP21], which we recall hereafter for the sake of completeness. First, recall that a function h on $\mathbb{R}^d$ is *blocked* if it can be written $h = \sum_{i=1}^{d^l} b_i \mathbf{1}_{A_i}$ where $b_1, \dots, b_{d^l}$ are non-negative real numbers and the A_i are axis-aligned rectangles in $\mathbb{R}^d$. Moreover, recall that the *Skorohod J_1-topology* on the space of càdlàg functions $D([0,T])$ is the topology induced by the metric:

$$d_{J_1}(f,g) := \inf_{\lambda}\left\{\|f\circ\lambda - g\|_\infty + \|\lambda - \mathrm{Id}_{[0,T]}\|_\infty\right\},$$

where the infimum is taken over all increasing continuous bijections of $[0,T]$.

Theorem V.16 (**Theorem 3.4 from [KRP21]**). *Let $T > 0$ and $X_1,\dots,X_n$ be sampled according to a bounded density g on $[0,1]^d$. Denote by $\mathcal{F}_n$ the Čech complex associated with the point cloud $n^{1/d}(X_i)_{i=1}^n$. Assume that blocked functions can uniformly approximate g. There is a Gaussian process $\mathfrak{G} : [0,T] \to \mathbb{R}$ such that for $t \in [0,T]$,*

$$\sqrt{n}\big(\chi_{\mathcal{F}_n}(t) - \mathbb{E}[\chi_{\mathcal{F}_n}(t)]\big) \underset{n\to\infty}{\longrightarrow} \mathfrak{G}(t),$$

in distribution in the Skorohod J_1-topology. Furthermore, there exist two real-valued functions γ and α such that the covariance of the limiting process is defined by:

$$\mathbb{E}[\mathfrak{G}(s)\mathfrak{G}(t)] = \mathbb{E}\left[\gamma\left(g(Z)^{1/d}(s,t)\right)\right] - \mathbb{E}\left[\alpha\left(g(Z)^{1/d}s\right)\right]\mathbb{E}\left[\alpha\left(g(Z)^{1/d}t\right)\right],$$

where Z is a random variable with density g.

We refer to [KRP21] for the expression of the two functions γ and α. Here again, the distribution of the points appears in the limiting object and, more precisely, in its covariance function. We can adapt this theorem to show that hybrid transforms of compactly supported kernels are also asymptotically normal.

Theorem V.17. *Consider the setting of Theorem V.16. Let $a, M > 0$ and $\kappa \in L^1(\mathbb{R})$. Further assume that κ is supported on $[0,T]$. Then, there is a Gaussian process $\tilde{\mathfrak{G}} : [a,M] \to \mathbb{R}$ such that:*

$$\sqrt{n}\left(\psi_{\mathcal{F}_n}^{\kappa} - \mathbb{E}\left[\psi_{\mathcal{F}_n}^{\kappa}\right]\right) \underset{n\to\infty}{\longrightarrow} \tilde{\mathfrak{G}} \quad a.s.,$$

in $\left(\mathcal{C}^0[a,M],\|\cdot\|_\infty\right)$. *Furthermore, the covariance of the limiting process is defined by:*

$$\mathbb{E}\left[\tilde{\mathfrak{G}}(\xi_1)\tilde{\mathfrak{G}}(\xi_2)\right] = \xi_1\xi_2 \int_0^{T/\xi_1}\int_0^{T/\xi_2} \kappa(\xi_1 t)\,\kappa(\xi_2 s)\,\operatorname{cov}\left(\mathfrak{G}(s),\mathfrak{G}(t)\right)\mathrm{d}s\,\mathrm{d}t,$$

where $\mathfrak{G}$ *is the Gaussian process defined in Theorem V.16.*

V.6.2 Limit theorem for multi-parameter hybrid transforms

Theorem V.18. *Assume that* Φ *is a stationary ergodic point process having finite moments. Let* $T, a > 0$ *and* $\kappa \in L^1(\mathbb{R})$. *Assume that* κ *is supported on* $[0,T]$. *We denote by* $\mathcal{F}_L$ *the filtration induced by the sublevel sets of* f *on* Φ_L. *Assume that there exists an increasing function* ρ *such that there exists* $i \in [\![1,m]\!]$ *such that for all* $(x,y) \in (\mathbb{R}^d)^2$,

$$\|x-y\| \leq \rho\left(f_i(\{x,y\})\right). \tag{V.9}$$

Under these assumptions, there exists a function $H : \mathbb{R}_+^{m*} \times \mathbb{R}_+ \to \mathbb{R}$ *that depends only on* κ *and* f *such that, for all* $\xi = (\xi_1,\ldots,\xi_m) \in \mathbb{R}_+^{m*}$ *and* $\lambda > a$,

$$\frac{1}{L^d}\psi^{\kappa}_{\mathcal{F}_L}(\lambda\xi) \underset{L\to\infty}{\longrightarrow} H(\xi,\lambda) \quad a.s..$$

This limit theorem is a direct consequence of the results from [HST18] for persistence diagrams of a large class of filtration functions. We refer to Section 3 of loc. cit. for the definition of a stationary ergodic point process. Note that this encompasses most cases of usual point processes such as Poisson, Ginibre, or Gibbs. This result makes use of the smoothness properties of the hybrid transforms and follows directly from Lemma V.9 that expresses restrictions of multi-parameter hybrid transforms to lines as one-parameter hybrid transforms. Similar results cannot be derived that easily for Euler characteristic profiles, as one would need to consider the joint law of several one-parameter filtrations. In addition, deriving a multi-dimensional central limit theorem from [PY01] would require the filter $\xi \cdot f$ to verify some translation invariance property. In practice, this very strong assumption is verified only by Čech and Vietoris-Rips filtrations as well as marked processes; see [BH22]. Alpha and function-Čech filtrations that we used in our experiments do not verify this assumption.

As pointed out in Example 1.3 of [HST18], Čech and Vietoris-Rips filtrations satisfy (V.9) for $\rho : t \mapsto 2t$. We provide below two examples of families a broad family of multi-parameter filtrations satisfying (V.9).

Example V.19. It is easy to check that the function-alpha filtration considered in the applications of Sections V.4.2 and V.4.3 satifies (V.9).

We give another class of filtrations satisfying (V.9) that contains in particular the distance-to-measure (DTM) filtrations; see [ACG+20].

Example V.20. Let h be a positive and bounded function from $\mathbb{R}^d$ to $\mathbb{R}$. The weighted Čech complex introduced in [ACG+20] is defined as follows. For every $x \in \mathbb{R}^d$ and real number $t \geq 0$, we define:

$$r_x(t) = \begin{cases} -\infty & \text{if } t < h(x), \\ t - h(x) & \text{otherwise.} \end{cases}$$

We denote by $\overline{B}_h(x,t) = \overline{B}\left(x, r_x(t)\right)$ the closed Euclidean ball of center x and radius $r_x(t)$. A simplex $\{x_0,\ldots,x_k\}$ in some finite set $\mathbb{X}$ belongs to the *weighted Čech complex* at scale $t \geq 0$ if the intersection of closed balls $\cap_{l=0}^k \overline{B}_h(x_l,t)$ is non-empty. Considering the weighted Čech complex for all scales t defines a filtration of $2^{\mathbb{X}}$ called *weighted Čech filtration*. The weighted Čech filtration satisfies (V.9) for $\rho : t \mapsto \max(\max h, 2t)$.

V.7 Proofs

In this section, we prove the results stated in Sections V.5 and V.6. In the following proofs, we make constant use of the fact that the distance $\widehat{d}_1$ may be computed on any decomposition of the functions and not only on minimal ones, that is, on signed barcodes. More precisely, for any decompositions $(\mathcal{C}^+, \mathcal{C}^-)$ and $(\mathcal{C}'^+, \mathcal{C}'^-)$ of two finitely presented functions φ and φ' respectively, one has:

$$\widehat{d}_1(\varphi, \varphi') = \inf\left\{\varepsilon > 0 : \exists \text{ bijection } h : \mathcal{C}^+ \cup \mathcal{C}'^- \to \mathcal{C}^- \cup \mathcal{C}'^+ \text{ with } \mathrm{cost}(h) \leq \varepsilon\right\}. \tag{V.10}$$

V.7.1 Proof of Proposition V.12

Recall that $m \geq 1$. Consider decompositions (η^+, η^-) and (η'^+, η'^-) of $\chi_{\mathcal{F}}$ and $\chi_{\mathcal{F}'}$ respectively. Assume there is a bijection $h : \eta^+ \cup \eta'^- \to \eta^- \cup \eta'^+$. If no such bijection exists, then $\widehat{d}_1(\chi_{\mathcal{F}}, \chi_{\mathcal{F}'})$ is infinite, and the inequality trivially holds. One has:

$$\chi_{\mathcal{F}} - \chi_{\mathcal{F}'} = \sum_{u \in \eta^+ \cup \eta'^-} \mathbf{1}_{Q_u} - \sum_{v \in \eta^- \cup \eta'^+} \mathbf{1}_{Q_v} = \sum_{u \in \eta^+ \cup \eta'^-} \mathbf{1}_{Q_u} - \mathbf{1}_{Q_{h(u)}}.$$

Therefore,

$$\|\chi_{\mathcal{F}} - \chi_{\mathcal{F}'}\|_{1,M} \leq \sum_{u \in \eta^+ \cup \eta'^-} \|\mathbf{1}_{Q_u} - \mathbf{1}_{Q_{h(u)}}\|_1. \tag{V.11}$$

By an elementary induction on $m \geq 1$, we can prove that for all $u, v \in \mathbb{R}^m$,

$$\|\mathbf{1}_{Q_u} - \mathbf{1}_{Q_v}\|_{1,M} \leq (2M)^{m-1}\|u - v\|_1.$$

This conludes the proof.

Assume now that $m = 1$. The existence of h ensures that $\|\chi_{\mathcal{F}} - \chi_{\mathcal{F}'}\|_1$ is finite and the result follows from (V.11) and the fact that $\|\mathbf{1}_{[u,v)}\|_1 = |u - v|$.

V.7.2 Proof of Corollary V.13

Let us prove the first inequality. Proposition V.12 with $m = 1$ ensures that for any $\xi \in K$,

$$\int_{\mathbb{R}} |\mathcal{R}_{\mathcal{F}}(\xi, s) - \mathcal{R}_{\mathcal{F}'}(\xi, s)| \, \mathrm{d}s = \|\xi_* \chi_{\mathcal{F}} - \xi_* \chi_{\mathcal{F}'}\|_1 \leq \widehat{d}_1(\xi_* \chi_{\mathcal{F}}, \xi_* \chi_{\mathcal{F}'}).$$

To prove the desired inequality, we will prove that $\widehat{d}_1(\xi_* \chi_{\mathcal{F}}, \xi_* \chi_{\mathcal{F}'}) \leq \|\xi\|_\infty \widehat{d}_1(\chi_{\mathcal{F}}, \chi_{\mathcal{F}'})$ for any $\xi \in \mathbb{R}_+^{m*}$. The result then follows from computing the q-norm on both sides. Consider decompositions (η^+, η^-) and (η'^+, η'^-) of $\chi_{\mathcal{F}}$ and $\chi_{\mathcal{F}'}$ respectively. They induce decompositions $(\xi_* \eta^+, \xi_* \eta^-)$ and $(\xi_* \eta'^+, \xi_* \eta'^-)$ of $\xi_* \chi_{\mathcal{F}} = \chi_{\xi_* \mathcal{F}}$ and $\xi_* \chi_{\mathcal{F}'} = \chi_{\xi_* \mathcal{F}'}$ respectively by the formula $\xi_* \eta^\pm = \{\xi \cdot u : u \in \eta^\pm\}$ and a similar one for $\mathcal{F}'$. Consider a bijection of multisets $h : \eta^+ \cup \eta'^- \to \eta^- \cup \eta'^+$. It induces a bijection of multisets $\xi_* h : \xi_* \eta^+ \cup \xi_* \eta'^- \to \xi_* \eta^- \cup \xi_* \eta'^+$ defined by $\xi \cdot u \mapsto \xi \cdot h(u)$ with cost:

$$\mathrm{cost}(\xi_* h) = \sum_{t \in \xi_* \eta^+ \cup \xi_* \eta'^-} \|t - \xi_* h(t)\|_1 = \sum_{u \in \eta^+ \cup \eta'^-} \|\xi \cdot u - \xi \cdot h(u)\|_1 \leq \|\xi\|_\infty \cdot \mathrm{cost}(h).$$

Taking the infimum over all bijections h yields $\widehat{d}_1(\xi_* \chi_{\mathcal{F}}, \xi_* \chi_{\mathcal{F}'}) \leq \|\xi\|_\infty \widehat{d}_1(\chi_{\mathcal{F}}, \chi_{\mathcal{F}'})$ by (V.10).

Let us now prove the second inequality. It follows from the definition of hybrid transforms that $\|\psi_{\mathcal{F}}^\kappa - \psi_{\mathcal{F}'}^\kappa\|_{L_K^q} \leq \|\kappa\|_\infty \|\mathcal{R}_{\mathcal{F}} - \mathcal{R}_{\mathcal{F}'}\|_{L_K^{q,1}}$ when κ is bounded. The first inequality yields the result.

V.7.3 Proof of Lemma V.14

The couple $\mathcal{C}_f = \big(\{f(\sigma)\}_{\dim\sigma \text{ even}}, \{f(\sigma)\}_{\dim\sigma \text{ odd}}\big)$ is a decomposition of χ_f. There is a similar decomposition $\mathcal{C}_g$ of χ_g. Moreover, the mapping $f(\sigma) \mapsto g(\sigma)$ induces a bijection of multisets $h : \mathcal{C}_f \to \mathcal{C}_g$ with $\mathrm{cost}(h) = \sum_{\sigma\in\mathcal{K}} \|f(\sigma) - g(\sigma)\|_1 = \|g - f\|_1$. The result follows from (V.10).

V.7.4 Proof of Theorem V.15

Let $X_1, \dots, X_n$ be an i.i.d. sample drawn according to an a.e. continuous bounded Lipschitz density g on $\mathbb{R}^d$. Consider a sequence $(r_n)_{n\in\mathbb{N}}$ such that $nr_n^d \to 0$ and for all $k \in [\![0, d-1]\!]$, $n^{k+2} r_n^{d(k+1)} \to \infty$ as $n \to \infty$.

Let us define $\Delta := \{(x,y) : 0 \le x \le y < \infty\} \cup \{(x,\infty) : 0 \le x < \infty\}$ and for every (s,t,u,v) such that $0 \le s \le t \le u \le v \le \infty$, denote by $R_{s,t,u,v}$ the rectangle $(s,t] \times (u,v]$ of Δ. Recall that, as done extensively in Section IV, a finite persistence diagram $D = \{(a_i, b_i)\}_{i=1,\dots,l}$ can be turned into a discrete measure $\mu = \sum_{i=1}^{l} \delta_{a_i,b_i}$ on Δ. Denote by $\mu_{k,n}$ the k-th persistence diagram of the Čech filtration of $1/r_n(X_i)_{i=1}^n$, seen as a discrete measure on Δ.

Theorem 3.2 of [Owa22] ensures that for every $k \in [\![0, d-1]\!]$ there exists a unique Radon measure μ_k on Δ such that we have the following vague convergence:

$$\frac{1}{n^{k+2} r_n^{d(k+1)}}\, \mu_{k,n} \xrightarrow[n\to\infty]{v} \frac{1}{(k+2)!} \left(\int_{\mathbb{R}^d} g^{k+2}(x)\, \mathrm{d}x \right) \mu_k \quad \text{a.s.}, \tag{V.12}$$

where for every $0 \le s \le t \le u \le v \le \infty$, there is an indicator geometric function $H_{s,t,u,v}$ on $\mathbb{R}^{d(k+2)}$ defined in [Owa22, Sec. 3.1], which does not depend on g and such that the measure μ_k is defined by:

$$\mu_k(R_{s,t,u,v}) = \int_{\mathbb{R}^{d(k+1)}} H_{s,t,u,v}(0, y_1, \dots, y_{k+1})\, \mathrm{d}y_1 \dots \mathrm{d}y_{k+1}.$$

Recall that the primitive kernel $\overline{\kappa}$ is such that $\overline{\kappa}(x) \to 0$ when $x \to +\infty$. Therefore, the fact that κ is supported on $[0,T]$ implies that the primitive $\overline{\kappa}$ is also supported on $[0,T]$. For $\xi > a$, denote by $h_\xi : (x,y) \in \Delta \mapsto \overline{\kappa}(\xi y) - \overline{\kappa}(\xi x)$. According to (V.3), one has:

$$\psi^{\kappa}_{\mathcal{F}_n}(\xi) = \sum_{k=0}^{d-1} (-1)^k \langle \mu_{k,n}, h_\xi \rangle.$$

Since h_ξ is continuous and supported on $[0, T/a]^2$, we have by the vague convergence in (V.12) that:

$$\frac{1}{n^{k+2} r_n^{d(k+1)}}\, \psi^{\kappa}_{\mathcal{F}_n}(\xi) \xrightarrow[n\to\infty]{} \sum_{k=0}^{d-1} \frac{(-1)^k}{(k+2)!} \left(\int_{\mathbb{R}^d} g^{k+2}(x) \mathrm{d}x \right) A_k(\xi) \quad \text{a.s.},$$

where $A_k(\xi) = \int_\Delta h_\xi \mathrm{d}\mu_k$.

V.7.5 Proof of Theorem V.17

Let $T > 0$ such that κ is supported in $[0,T]$. Let $a, M > 0$ and let $\xi \in [a, M]$. According to (V.7), we have that:

$$\psi^{\kappa}_{\mathcal{F}}(\xi) = \xi \int_0^{T/\xi} \kappa(\xi \cdot t) \chi_{\mathcal{F}}(t) \mathrm{d}t,$$

and similarly for $\chi_{\mathcal{F}_n}$. Since κ is in L^1, the mappings $\psi^\kappa_{\mathcal{F}}$ and $\psi^\kappa_{\mathcal{F}_n}$ are continuous on $[a, M]$. According to Theorem V.16, there is a Gaussian process $\mathfrak{G} : [0, T/a] \to \mathbb{R}_+$ such that for all $t \in [0, T/a]$, we have that:

$$\sqrt{n}\left(\chi_{\mathcal{F}_n}(t) - \mathbb{E}[\chi_{\mathcal{F}_n}(t)]\right) \underset{n\to\infty}{\longrightarrow} \mathfrak{G}(t), \tag{V.13}$$

in distribution in the Skorohod J_1-topology. Therefore, by linearity of the mapping $\chi \mapsto \xi \int_0^{T/\xi} \kappa(\xi \cdot t)\chi(t)\mathrm{d}t$, we have that:

$$\sqrt{n}\left(\psi^\kappa_{\mathcal{F}_n} - \mathbb{E}\left[\psi^\kappa_{\mathcal{F}_n}\right]\right) \;=\; \xi \int_0^{T/\xi} \kappa(\xi \cdot t)\left[\sqrt{n}\big(\chi_{\mathcal{F}_n}(t) - \mathbb{E}[\chi_{\mathcal{F}_n}(t)]\big)\right] \mathrm{d}t$$

Denote by φ the mapping from the space of càdlàg functions $D([0, T])$ with Skorohod J_1-topology to $(\mathcal{C}^0([a, M]), \|\cdot\|_\infty)$ defined by:

$$\varphi : \chi \mapsto \left(\xi \mapsto \xi \int_0^{T/\xi} \kappa(\xi \cdot t)\chi(t)\mathrm{d}t\right).$$

We, therefore, have that:

$$\sqrt{n}\left(\psi^\kappa_{\mathcal{F}_n} - \mathbb{E}\left[\psi^\kappa_{\mathcal{F}_n}\right]\right) \;=\; \varphi\left(\sqrt{n}\left(\chi_{\mathcal{F}_n} - \mathbb{E}[\chi_{\mathcal{F}_n}]\right)\right).$$

It is easy to check that:

$$\|\varphi(\chi_1) - \varphi(\chi_2)\|_\infty \leq \frac{M}{a}\|\chi_1 - \chi_2\|_\infty \int_0^T |\kappa(u)|\mathrm{d}u,$$

so that the mapping φ is Lipschitz and, therefore, continuous. Thus, the continuous mapping theorem along with (V.13) yields that almost surely, one has the following convergence in $(\mathcal{C}^0([a, M]), \|\cdot\|_\infty)$,

$$\sqrt{n}\left(\psi^\kappa_{\mathcal{F}_n} - \mathbb{E}\left[\psi^\kappa_{\mathcal{F}_n}\right]\right) \underset{n\to\infty}{\longrightarrow} \tilde{\mathfrak{G}}(\xi) := \xi \int_0^{T/\xi} \kappa(\xi \cdot t)\mathfrak{G}(t)\,\mathrm{d}t.$$

The covariance of the limiting process $\tilde{\mathfrak{G}}$ follows immediately from that of $\mathfrak{G}$.

V.7.6 Proof of Theorem V.18

Let $\xi = (\xi_1, \ldots, \xi_m) \in \mathbb{R}_+^{m*}$. Denote by $\mu_{k,L}^{\xi_*\mathcal{F}}$ the measure associated with the k-th persistence diagram of Φ_L for the filtration function $\xi \cdot f = \sum_{i=1}^m \xi_i f_i$. By hypothesis, there exists $i \in [\![1, m]\!]$ such that for all $(x, y) \in (\mathbb{R}^d)^2$, $\|x - y\| \leq \rho\left(f_i(\{x, y\})\right)$. Let $\rho' : x \mapsto \rho(x/\xi_i)$. Therefore, as the filtration functions are non-negative and ρ and ρ' are increasing, we have that:

$$\rho'\left(\sum_{j=1}^m \xi_j f_j(\{x, y\})\right) \geq \rho'(\xi_i f_i(\{x, y\}) \geq \rho(f_i(\{x, y\})) \geq \|x - y\|. \tag{V.14}$$

The filtration function $\xi \cdot f$ therefore verifies all the hypotheses of the Theorem 1.5 of [HST18], which states that there exists a Radon measure ν_k such that almost surely, we have the vague convergence $\frac{1}{L^d}\mu_{k,L}^{\xi_*\mathcal{F}} \xrightarrow{v} \nu_k^{\xi\cdot f}$ as $L \to \infty$. Note that in loc. cit., the authors make the additional hypothesis that the filtration function is translation invariant. However, this assumption is

only needed to derive a central limit theorem on persistent Betti numbers but not required for the above law of large numbers, for which we only need (V.14) to hold. As in the proof of Theorem V.15, we introduce a continuous function $h_\lambda : (x, y) \in \Delta \mapsto \overline{\kappa}(\lambda y) - \overline{\kappa}(\lambda x)$. This function is supported on $[0, T/a]^2$. According to (V.3) together with Lemma V.9, we have that:

$$\psi^{\kappa}_{\mathcal{F}_L}(\lambda\xi) = \sum_{k=0}^{d-1} (-1)^k \langle \mu^{\xi_*\mathcal{F}}_{k,L}, h_\lambda \rangle.$$

Hence the result, by the vague convergence $\frac{1}{L^d}\mu^{\xi_*\mathcal{F}}_{k,L} \xrightarrow{v} \nu^{\xi\cdot f}_k$ for every $k \in [\![0, d-1]\!]$.

VI Conclusion, future directions

In this thesis, we studied persistence diagrams, their uses in machine learning, and their limitations. In Section III, we demonstrated that the total persistence for the sublevel sets filtration is a strong regularizer and can help reconstruct a smooth function from noisy observations. We have demonstrated some theoretical guarantees, especially Theorem III.6. This theorem gives an oracle inequality assuming that the function we try to estimate lies in the span of a finite number of Laplace eigenfunctions. Two possible extension directions of this result would be to compute the bias term where the regression function is assumed to be square-integrable and obtain a similar guarantee when we estimate the regression function using the eigenvectors of the graph Laplacian. On the numerical side, this procedure is costly because it requires computing many persistence diagrams, and speeding up the optimization remains an open research question. Furthermore, this technical limitation constrained us is considering only total persistence. However, other statistics computing the distance of points to the diagonal, including weights over all homological dimensions, should be investigated.

We have then demonstrated in Section IV that persistence diagrams contain relevant information to classify data and extract qualitative information about their topology. In addition, persistence diagrams of random complexes are an active research topic. We have proposed the Theorem IV.13 and Corollary IV.14 where we study the asymptotic of persistence diagrams of the Čech complex of a random sample in a particular sampling regime called the sparse regime. Extensions to different regimes and deriving finer concentration inequalities remain open questions.

Finally, we showed in Section V that persistence diagrams are less efficient than simpler descriptors based on Euler characteristic computations. Indeed, we have introduced two new descriptors, the Radon and hybrid transforms, and thoroughly studied an already existing one: the Euler characteristic profile. We have proven that these descriptors are much faster to compute than persistence diagrams, typically having better accuracy and similar explainability properties. Furthermore, these descriptors naturally generalize to multi-parameter persistence, which is not permitted by persistence diagrams. A take-home message is that when tackling a classification problem where the data is believed to contain some relevant topological information, Euler-based descriptors must be preferred to persistence diagrams due to their simplicity and overall high accuracy. The study in this manuscript is limited to simplicial complexes, and this work should be extended to cubical complexes. In this case, many possible filtration functions can be considered, and a thorough comparison with state-of-the-art methods in image processing would be a consequent line of research. In addition, these tools allow for non-increasing functions on simplicial complexes, which could extend to the analysis of time-varying complexes.

References

[AAPL22] Alberto Alcalá-Alvarez and Pablo Padilla-Longoria. A framework for topological music analysis (tma). arXiv preprint arXiv:2204.09744, 2022.

[ACC+21] Andrew Aukerman, Mathieu Carrière, Chao Chen, Kevin Gardner, Raúl Rabadán, and Rami Vanguri. Persistent homology based characterization of the breast cancer immune microenvironment: a feasibility study. Journal of Computational Geometry, 12(2):183–206, 2021.

[ACG+20] Hirokazu Anai, Frédéric Chazal, Marc Glisse, Yuichi Ike, Hiroya Inakoshi, Raphaël Tinarrage, and Yuhei Umeda. Dtm-based filtrations. In Topological Data Analysis: The Abel Symposium 2018, pages 33–66. Springer, 2020.

[AEK+17] 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.

[AFH+13] 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.

[AL19] Eddie Aamari and Clément Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. Ann. Statist., 47(1):177–204, 2019.

[Amo13] Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. Artificial intelligence, 201:81–105, 2013.

[AQO+22] Erik J Amézquita, Michelle Y Quigley, Tim Ophelders, Jacob B Landis, Daniel Koenig, Elizabeth Munch, and Daniel H Chitwood. Measuring hidden phenotype: Quantifying the shape of barley seeds using the Euler Characteristic Transform. in silico Plants, 4(1), 2022.

[AT09] Robert J. Adler and Jonathan E. Taylor. Random fields and geometry. Springer Science & Business Media, 2009.

[B+15] Peter Bubenik et al. Statistical topological data analysis using persistence landscapes. J. Mach. Learn. Res., 16(1):77–102, 2015.

[BA14] Omer Bobrowski and Robert Adler. Distance functions, critical points, and the topology of random Cech complexes. Homology, Homotopy and Applications, 16(2):311–344, 2014.

[BB21] Francisco J Baldán and José M Benítez. Multivariate times series classification through an interpretable representation. Information Sciences, 569:596–614, 2021.

[BBM99] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. Probability Theory and Related Fields, 113(3):301–413, 1999.

[BCY18] Jean-Daniel Boissonnat, Frédéric Chazal, and Mariette Yvinec. Geometric and topological inference, volume 57. Cambridge University Press, 2018.

[BDL+18] 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. arXiv preprint arXiv:1811.00075, 2018.

[BE17] Ulrich Bauer and Herbert Edelsbrunner. The Morse theory of Čech and Delaunay complexes. Transactions of the American Mathematical Society, 369(5):3741–3762, 2017.

[BGNDS20] Rickard Brüel-Gabrielsson, Bradley J Nelson, Anjan Dwaraknath, and Primoz Skraba. A topology layer for machine learning. In International Conference on Artificial Intelligence and Statistics, pages 1553–1563. PMLR, 2020.

[BH22] Magnus B Botnan and Christian Hirsch. On the consistency and asymptotic normality of multiparameter persistent Betti numbers. Journal of Applied and Computational Topology, pages 1–38, 2022.

[BHHS22] Clément Berenfeld, John Harvey, Marc Hoffmann, and Krishnan Shankar. Estimating the reach of a manifold via its convexity defect function. Discrete & Computational Geometry, 67(2):403–438, 2022.

[BHPW20] Peter Bubenik, Michael Hull, Dhruv Patel, and Benjamin Whittle. Persistent homology detects curvature. Inverse Problems, 36(2):025008, 2020.

[BIK15] Dmitri Burago, Sergei Ivanov, and Yaroslav Kurylev. A graph discretization of the Laplace–Beltrami operator. Journal of Spectral Theory, 4(4):675–714, 2015.

[BK18] Omer Bobrowski and Matthew Kahle. Topology of random geometric complexes: a survey. Journal of applied and Computational Topology, 1(3):331–364, 2018.

[BLW96] Peter L Bartlett, Philip M Long, and Robert C Williamson. Fat-shattering and the learnability of real-valued functions. Journal of computer and system sciences, 52(3):434–452, 1996.

[BM14] Jean-Daniel Boissonnat and Clément Maria. The simplex tree: An efficient data structure for general simplicial complexes. Algorithmica, 70:406–427, 2014.

[BM15] Omer Bobrowski and Sayan Mukherjee. The topology of probability distributions on manifolds. Probability theory and related fields, 161(3):651–686, 2015.

[BMT17] Omer Bobrowski, Sayan Mukherjee, and Jonathan E. Taylor. Topological consistency via kernel estimation. Bernoulli, 23(1):288 – 328, 2017.

[BMTY05] M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. International Journal of Computer Vision, 61(2):139–157, 2005.

[BN03] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 15(6):1373–1396, 2003.

[BNS06] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, 7(85):2399–2434, 2006.

[BSA+22] Gabriele Beltramo, Primoz Skraba, Rayna Andreeva, Rik Sarkar, Ylenia Giarratano, and Miguel O Bernabeu. Euler characteristic surfaces. Foundations of Data Science, 4(4):505–536, 2022.

[Bv11] Peter Bühlmann and Sara van de Geer. Statistics for High-dimensional Data: Methods, Theory and Applications. Springer Science & Business Media, 2011.

[BW20] Peter Bubenik and Alexander Wagner. Embeddings of persistence diagrams into Hilbert spaces. Journal of Applied and Computational Topology, 4(3):339–351, 2020.

[CB18] Mathieu Carrière and Ulrich Bauer. On the Metric Distortion of Embedding Persistence Diagrams into Separable Hilbert Spaces. In Proceedings of the thirty-fifth International Symposium on Computational Geometry, 2018.

[CB20] Mathieu Carrière and Andrew Blumberg. Multiparameter persistence image for topological machine learning. Advances in Neural Information Processing Systems, 33:22432–22444, 2020.

[CCG+21] Mathieu Carriere, Frédéric Chazal, Marc Glisse, Yuichi Ike, Hariprasad Kannan, and Yuhei Umeda. Optimizing persistent homology based functions. In International Conference on Machine Learning, pages 1294–1303. PMLR, 2021.

[CCI+20] 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.

[CCO17] Mathieu Carriere, Marco Cuturi, and Steve Oudot. Sliced Wasserstein kernel for persistence diagrams. In International conference on machine learning, pages 664–673. PMLR, 2017.

[CDSO14] Frédéric Chazal, Vin De Silva, and Steve Oudot. Persistence stability for geometric complexes. Geometriae Dedicata, 173(1):193–214, 2014.

[CG16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794, 2016.

[CGLM14] 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, 2014.

[CGLS16] Yongwan Chun, Daniel A Griffith, Monghyeon Lee, and Parmanand Sinha. Eigenvector selection with stepwise regression techniques to construct eigenvector spatial filters. Journal of Geographical Systems, 18(1):67–85, 2016.

[CGTL21] Jeff Calder, Nicolás García Trillos, and Marta Lewicka. Lipschitz regularity of graph Laplacians on random data clouds. SIAM Journal on Mathematical Analysis (to appear), 2021.

[Cha84] Isaac Chavel. Eigenvalues in Riemannian Geometry. Academic Press, 1984.

[Cha23] Frédéric Chazal. Persistence for TDA. https://geometrica.saclay.inria.fr/team/Fred.Chazal/slides/PersistenceForTDA_2023.pdf, 2023.

[Chu97] Fan RK Chung. Spectral graph theory, volume 92. American Mathematical Soc., 1997.

[CL06] Ronald R Coifman and Stéphane Lafon. Diffusion maps. Applied and Computational Harmonic Analysis, 21(1):5–30, 2006.

[CLR21] Frédéric Chazal, Clément Levrard, and Martin Royer. Clustering of measures via mean measure quantization. Electronic Journal of Statistics, 15(1):2060–2104, 2021.

[CM21] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. Frontiers in artificial intelligence, 4:667963, 2021.

[CMT22] Justin Curry, Sayan Mukherjee, and Katharine Turner. How many directions determine a shape and other sufficiency results for two topological transforms. Transactions of the American Mathematical Society, Series B, 9(32):1006–1043, 2022.

[CNBW19] Chao Chen, Xiuyan Ni, Qinxun Bai, and Yusu Wang. A topological regularizer for classifiers via persistent homology. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 2573–2582, 2019.

[CSEH07] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. Discrete & Computational Geometry, 37(1):103–120, 2007.

[CSEM06] David Cohen-Steiner, Herbert Edelsbrunner, and Dmitriy Morozov. Vines and vineyards by updating persistence in linear time. In Proceedings of the twenty-second annual symposium on Computational geometry, pages 119–126, 2006.

[CZ09] Gunnar Carlsson and Afra Zomorodian. The theory of multidimensional persistence. Discrete & Computational Geometry, 42(1):71–93, 2009.

[CZCG04] Gunnar Carlsson, Afra Zomorodian, Anne Collins, and Leonidas Guibas. Persistence barcodes for shapes. In Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing, pages 124–135, 2004.

[DDQHD13] 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.

[DG22] Paweł Dłotko and Davide Gurnari. Euler characteristic curves and profiles: a stable shape invariant for big data problems. arXiv preprint:2212.01666, 2022.

[DHS+13] 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.

[DL21] Vincent Divol and Théo Lacombe. Understanding the topology and the geometry of the space of persistence diagrams via optimal partial transport. Journal of Applied and Computational Topology, 5:1–53, 2021.

[DLLP97] 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.

[DM21] Laurent Decreusefond and Guillaume Moroz. Optimal transport between determinantal point processes and application to fast simulation. Modern Stochastics: Theory and Applications, 8(2):209–237, 2021.

[DN18] Prerona Dutta and Khai T Nguyen. Covering numbers for bounded variation functions. Journal of Mathematical Analysis and Applications, 468(2):1131–1143, 2018.

[DP19] Vincent Divol and Wolfgang Polonik. On the choice of weight functions for linear representations of persistence diagrams. Journal of Applied and Computational Topology, 3(3):249–283, 2019.

[DWW21] David B Dunson, Hau-Tieng Wu, and Nan Wu. Spectral convergence of graph Laplacian and heat kernel reconstruction in ℓ_∞ from random samples. Applied and Computational Harmonic Analysis, 2021.

[EH22] Herbert Edelsbrunner and John L Harer. Computational topology: an introduction. American Mathematical Society, 2022.

[ELZ00] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. In Proceedings 41st annual symposium on foundations of computer science, pages 454–463. IEEE, 2000.

[Fed59] Herbert Federer. Curvature measures. Trans. Amer. Math. Soc., 93:418–491, 1959.

[FL01] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96:1348–1360, 2001.

[FM22] Ximena Fernández and Diego Mateos. Topological biomarkers for real-time detection of epileptic seizures. arXiv preprint arXiv:2211.02523, 2022.

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

[FV06] Frédéric Ferraty and Philippe Vieu. Nonparametric functional data analysis: theory and practice, volume 76. Springer, 2006.

[GBL18] Franziska Göbel, Gilles Blanchard, and Ulrike von Luxburg. Construction of tight frames on graphs and application to denoising. In W. K. Härdle, H. H-S. Lu, and X. Shen, editors, Handbook of Big Data Analytics, Springer Handbooks of Computational Statistics, pages 503–522. 2018.

[GH21] Dejan Govc and Richard Hepworth. Persistent magnitude. Journal of Pure and Applied Algebra, 225(3), 2021.

[Gir14] Christophe Giraud. Introduction to High-Dimensional Statistics, volume 138. CRC Press, 2014.

[GLM18] Robert Ghrist, Rachel Levanger, and Huy Mai. Persistent homology and Euler integral transforms. Journal of Applied and Computational Topology, 2:55–60, 2018.

[GR11] Robert Ghrist and Michael Robinson. Euler–Bessel and Euler–Fourier transforms. Inverse Problems, 27(12), 2011.

[Gri09] Alexander Grigoryan. Heat Kernel and Analysis on Manifolds, volume 47 of AMS/IP Studies in Advanced Mathematics. American Mathematical Soc., 2009.

[GTGHS20] Nicolás García Trillos, Moritz Gerlach, Matthias Hein, and Dejan Slepčev. Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace–Beltrami operator. Foundations of Computational Mathematics, 20(4):827–887, 2020.

[GWR+14] Ricardo Guerrero, Robin Wolz, A W Rao, Daniel Rueckert, and Alzheimer's Disease Neuroimaging Initiative (ADNI). Manifold population modeling as a neuro-imaging biomarker: application to ADNI and ADNI-GO. NeuroImage, 94:275–286, 2014.

[HB05] Matthias Hein and Olivier Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In International Workshop on Artificial Intelligence and Statistics, pages 136–143. PMLR, 2005.

[HBB+22] Olympio Hacquard, Krishnakumar Balasubramanian, Gilles Blanchard, Clément Levrard, and Wolfgang Polonik. Topologically penalized regression on manifolds. Journal of Machine Learning Research, 23(161):1–39, 2022.

[HBL23] Olympio Hacquard, Gilles Blanchard, and Clément Levrard. Statistical learning on measures: an application to persistence diagrams. arXiv preprint arXiv:2303.08456, 2023.

[Hen90] Harrie Hendriks. Nonparametric estimation of a probability density on a Riemannian manifold using Fourier expansions. The Annals of Statistics, pages 832–849, 1990.

[Hes20] Kathryn Hess. Topological adventures in neuroscience. In Topological Data Analysis: The Abel Symposium 2018, pages 277–305. Springer, 2020.

[HK70] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1):55–67, 1970.

[HL23] Olympio Hacquard and Vadim Lebovici. Euler characteristic tools for topological data analysis. arXiv preprint arXiv:2303.14040, 2023.

[HMR21] Felix Hensel, Michael Moor, and Bastian Rieck. A survey of topological machine learning methods. Frontiers in Artificial Intelligence, 4:681108, 2021.

[HNH+16] Yasuaki Hiraoka, Takenobu Nakamura, Akihiko Hirata, Emerson G Escolar, Kaname Matsue, and Yasumasa Nishiura. Hierarchical structures of amorphous solids characterized by persistent homology. Proceedings of the National Academy of Sciences, 113(26):7035–7040, 2016.

[HR16] Jan-Christian Hütter and Philippe Rigollet. Optimal rates for total variation denoising. In Conference on Learning Theory, pages 1115–1146, 2016.

[HST18] Yasuaki Hiraoka, Tomoyuki Shirai, and Khanh Duy Trinh. Limit theorems for persistence diagrams. The Annals of Applied Probability, 28(5):2740–2780, 2018.

[HW17] Teresa Heiss and Hubert Wagner. Streaming algorithm for Euler characteristic curves of multidimensional images. In Computer Analysis of Images and Patterns: 17th International Conference, CAIP 2017, Ystad, Sweden, August 22-24, 2017, Proceedings, Part I 17, pages 397–409. Springer, 2017.

[IOH20] Takashi Ichinomiya, Ippei Obayashi, and Yasuaki Hiraoka. Protein-folding analysis using features obtained by persistent homology. Biophysical Journal, 118(12):2926–2937, 2020.

[IS14] Jeff Irion and Naoki Saito. Hierarchical graph Laplacian eigen-transforms. JSIAM Letters, 6:21–24, 2014.

[Ivr16] Victor Ivrii. 100 years of Weyl's law. Bulletin of Mathematical Sciences, 6(3):379–452, 2016.

[JKN20] Qitong Jiang, Sebastian Kurtek, and Tom Needham. The weighted Euler curve transform for shape and image analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 844–845, 2020.

[JZC+15] Biao Jie, Daoqiang Zhang, Bo Cheng, Dinggang Shen, and Alzheimer's Disease Neuroimaging Initiative (ADNI). Manifold regularized multitask feature learning for multimodality disease classification. Human Brain Mapping, 36(2):489–507, 2015.

[KDS+18] Lida Kanari, Paweł Dłotko, Martina Scolamiero, Ran Levi, Julian Shillcock, Kathryn Hess, and Henry Markram. A topological representation of branching neuronal morphologies. Neuroinformatics, 16:3–13, 2018.

[KHF16] 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.

[KKCM23] 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.

[KMK+20] Tomáš Kocák, Rémi Munos, Branislav Kveton, Shipra Agrawal, and Michal Valko. Spectral bandits. 2020.

[Kol98] Vladimir I. Koltchinskii. Asymptotics of spectral projections of some random matrices approximating integral operators. In Ernst Eberlein, Marjorie Hahn, and Michel Talagrand, editors, High Dimensional Probability, pages 191–227, Basel, 1998. Birkhäuser Basel.

[Kon14] Aryeh Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In International Conference on Machine Learning, pages 28–36. PMLR, 2014.

[KRP21] Johannes Krebs, Benjamin Roycraft, and Wolfgang Polonik. On approximation theorems for the Euler characteristic with applications to the bootstrap. Electronic Journal of Statistics, 15(2):4462–4509, 2021.

[KS94] Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. Journal of Computer and System Science, 48:464–497, 1994.

[LBD+17] Yongjin Lee, Senja D Barthel, Paweł Dłotko, S Mohamad Moosavi, Kathryn Hess, and Berend Smit. Quantifying similarity of pore-geometry in nanoporous materials. Nature communications, 8(1):1–8, 2017.

[LCB22] David Loiseaux, Mathieu Carriere, and Andrew J Blumberg. Efficient approximation of multiparameter persistence modules. arXiv preprint arXiv:2206.02026, 2022.

[LCS22] David Loiseaux, Mathieu Carrière, and Hannah Schreiber. Multipersistence Modules Approximation (`MMA`). https://github.com/DavidLapous/multipers, 2022.

[Leb22] Vadim Lebovici. Hybrid transforms of constructible functions. Foundations of Computational Mathematics, pages 1–47, 2022.

[Lee90] A. J. Lee. U-statistics, volume 110 of Statistics: Textbooks and Monographs. Marcel Dekker, Inc., New York, 1990. Theory and practice.

[Lei13] Tom Leinster. The magnitude of metric spaces. Documenta Mathematica, 18:857–905, 2013.

[Lev06] Bruno Levy. Laplace-Beltrami eigenfunctions towards an algorithm that "understands" geometry. In IEEE International Conference on Shape Modeling and Applications 2006 (SMI'06), pages 13–13. IEEE, 2006.

[LW16] Michael Lesnick and Matthew Wright. Interactive visualization of 2-D persistence Modules. 2016.

[LY18] Tam Le and Makoto Yamada. Persistence fisher kernel: A riemannian manifold kernel for persistence diagrams. Advances in Neural Information Processing Systems, 31, 2018.

[Mas07] Pascal Massart. Concentration Inequalities and Model Selection. Springer, 2007.

[Mau16] Andreas Maurer. A chain rule for the expected suprema of Gaussian processes. Theoretical Computer Science, 650:109–122, 2016.

[MBGY14] 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.

[MBP22] Martín Mijangos, Alessandro Bravetti, and Pablo Padilla. Musical stylistic analysis: A study of intervallic transition graphs via persistent homology. arXiv preprint arXiv:2204.11139, 2022.

[MC20] Caroline Moosmüller and Alexander Cloninger. Linear Optimal Transport embedding: provable Wasserstein classification for certain rigid transformations and perturbations. arXiv preprint arXiv:2008.09165, 2020.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[MFDS12] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. Advances in neural information processing systems, 25, 2012.

[Mil63] John Milnor. Morse theory. Based on lecture notes by M. Spivak and R. Wells., volume 51 of Annals of Mathematics Studies. Princeton University Press, 1963.

[MM07] Sridhar Mahadevan and Mauro Maggioni. Proto-value Functions: A Laplacian Framework for Learning Representation and Control in Markov Decision Processes. Journal of Machine Learning Research, 8(10), 2007.

[MMS11] Nikola Milosavljević, Dmitriy Morozov, and Primoz Skraba. Zigzag persistent homology in matrix multiplication time. In Proceedings of the twenty-seventh Annual Symposium on Computational Geometry, pages 216–225, 2011.

[Moh91] Bojan Mohar. The Laplacian spectrum of graphs. Graph Theory, Combinatorics, and Applications, 2(871-898):12, 1991.

[MRT18] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of Machine Learning. MIT press, 2018.

[MV03] Shahar Mendelson and Roman Vershynin. Entropy and the combinatorial dimension. Inventiones mathematicae, 152(1):37–55, 2003.

[NJW02] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems, 2:849–856, 2002.

[NSJ07] Jens Nilsson, Fei Sha, and Michael I Jordan. Regression on manifolds using kernel dimension reduction. In Proceedings of the 24th International Conference on Machine learning, pages 697–704, 2007.

[OHK18] 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.

[OS21] Steve Oudot and Luis Scoccola. On the stability of multigraded Betti numbers and Hilbert functions. arXiv preprint:2112.11901, 2021.

[Ott22] Nina Otter. Magnitude meets persistence: homology theories for filtered simplicial sets. Homology, Homotopy and Applications, 24(2):365–387, 2022.

[Owa22] Takashi Owada. Convergence of persistence diagram in the sparse regime. The Annals of Applied Probability, 32(6):4706–4736, 2022.

[PEVdW+17] Pratyush Pranav, Herbert Edelsbrunner, Rien Van de Weygaert, Gert Vegter, Michael Kerber, Bernard JT Jones, and Mathijs Wintraecken. The topology of the cosmic web in terms of persistent Betti numbers. Monthly Notices of the Royal Astronomical Society, 465(4):4281–4310, 2017.

[PPS19] Iosif Polterovich, Leonid Polterovich, and Vukašin Stojisavljević. Persistence barcodes and Laplace eigenfunctions on surfaces. Geometriae Dedicata, 201(1):111–138, 2019.

[PRSZ20] Leonid Polterovich, Daniel Rosen, Karina Samvelyan, and Jun Zhang. Topological persistence in geometry and analysis, volume 74. American Mathematical Society, 2020.

[PSRW13] Barnabás Póczos, Aarti Singh, Alessandro Rinaldo, and Larry Wasserman. Distribution-free distribution regression. In Artificial Intelligence and Statistics, pages 507–515. PMLR, 2013.

[PVG+11] 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.

[PY01] Mathew D Penrose and Joseph E Yukich. Central limit theorems for some graphs in computational geometry. Annals of Applied probability, pages 1005–1041, 2001.

[RB19] Raúl Rabadán and Andrew J Blumberg. Topological data analysis for genomics and evolution: topology in biology. Cambridge University Press, 2019.

[RCB21] Raphael Reinauer, Matteo Caorsi, and Nicolas Berkouk. Persformer: A transformer architecture for topological machine learning. arXiv preprint arXiv:2112.15210, 2021.

[RCL+21] 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.

[RHBK15] 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.

[ROF92] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena, 60(1-4):259–268, 1992.

[Ros97] Steven Rosenberg. The Laplacian on a Riemannian manifold: An introduction to analysis on manifolds. Number 31. Cambridge University Press, 1997.

[Rou15] Vincent Rouvreau. Alpha complex. In GUDHI User and Reference Manual. GUDHI Editorial Board, 2015.

[RSL20] Bastian Rieck, Filip Sadlo, and Heike Leitte. Topological machine learning with persistence indicator functions. In Topological Methods in Data Analysis and Visualization V: Theory, Algorithms, and Applications 7, pages 87–101. Springer, 2020.

[RYB+20] Bastian Rieck, Tristan Yates, Christian Bock, Karsten Borgwardt, Guy Wolf, Nicholas Turk-Browne, and Smita Krishnaswamy. Uncovering the topology of time-varying fmri data using cubical persistence. Advances in neural information processing systems, 33:6900–6912, 2020.

[Sai08] Naoki Saito. Data analysis and representation on a general domain using eigenfunctions of Laplacian. Applied and Computational Harmonic Analysis, 25(1):68–97, 2008.

[Sch89] Pierre Schapira. Cycles lagrangiens, fonctions constructibles et applications. Séminaire Équations aux dérivées partielles (Polytechnique) dit aussi "Séminaire Goulaouic-Schwartz", 1988-1989.

[Sch95] Pierre Schapira. Tomography of constructible functions. In International Symposium on Applied Algebra, Algebraic Algorithms, and Error-Correcting Codes, pages 427–435. Springer, 1995.

[Sim97] Hans Ulrich Simon. Bounds on the number of examples needed for learning functions. SIAM Journal on Computing, 26(3):751–763, 1997.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.

[SSG+18] Areejit Samal, RP Sreejith, Jiao Gu, Shiping Liu, Emil Saucan, and Jürgen Jost. Comparative analysis of two discretizations of ricci curvature for complex networks. Scientific reports, 8(1):8650, 2018.

[SSPG16] 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.

[ST12] Sivan Sabato and Naftali Tishby. Multi-instance learning with any hypothesis class. The Journal of Machine Learning Research, 13(1):2999–3039, 2012.

[ST20] Primoz Skraba and Katharine Turner. Wasserstein stability for persistence diagrams. arXiv preprint arXiv:2006.16824, 2020.

[SX10] Yiqian Shi and Bin Xu. Gradient estimate of an eigenfunction on a compact Riemannian manifold without boundary. Annals of Global Analysis and Geometry, 38(1):21–26, 2010.

[SZ21] Alexander Smith and Victor M Zavala. The Euler characteristic: A general topological descriptor for complex data. Computers & Chemical Engineering, 154:107463, 2021.

[Tal03] Michel Talagrand. Vapnik–Chervonenkis type conditions and uniform Donsker classes of functions. The Annals of Probability, 31(3):1565–1582, 2003.

[Tro15] Joel A Tropp. An introduction to matrix concentration inequalities. Foundations and Trends in Machine Learning, 8(1-2):1–230, 2015.

[TVH19] Quoc Hoan Tran, Van Tuan Vo, and Yoshihiko Hasegawa. Scale-variant topological information for characterizing the structure of complex networks. Phys. Rev. E, 100:032308, 2019.

[vdGB09] Sara A van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. Electronic Journal of Statistics, 3:1360–1392, 2009.

[VdMH08] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. Journal of machine learning research, 9(11), 2008.

[Ver18] Roman Vershynin. High-Dimensional Probability: An Introduction with Applications in Data Science, volume 47 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[Vip20] Oliver Vipond. Multiparameter persistence landscapes. Journal of Machine Learning Research, 21(61):1–38, 2020.

[Vir88] Oleg Yanovich Viro. Some integral calculus based on Euler characteristic. In Topology and geometry—Rohlin seminar, pages 127–138. Springer, 1988.

[vLBB08] Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. The Annals of Statistics, pages 555–586, 2008.

[Vov13] Vladimir Vovk. Kernel ridge regression. Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik, pages 105–116, 2013.

[VZ17] Saurabh Verma and Zhi-Li Zhang. Hunt for the unique, stable, sparse and fast feature learning on graphs. Advances in Neural Information Processing Systems, 30, 2017.

[Wik23] Wikipedia. Čech complex — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=%C4%8Cech%20complex&oldid=1149764739, 2023. [Online; accessed 15-May-2023].

[WSST15] Yu-Xiang Wang, James Sharpnack, Alex Smola, and Ryan Tibshirani. Trend filtering on graphs. In Artificial Intelligence and Statistics, pages 1042–1050. PMLR, 2015.

[WYHY15] 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.

[XATZ22] Lu Xian, Henry Adams, Chad M Topaz, and Lori Ziegelmeier. Capturing dynamics of time-varying data via topology. Foundations of Data Science, 4(1):1–36, 2022.

[XHLJ19] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In International Conference on Learning Representations, 2019.

[YSA17] 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.

[ZC04] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. In Proceedings of the twentieth annual symposium on Computational geometry, pages 347–356, 2004.

[Zel08] Steve Zelditch. Local and global analysis of eigenfunctions. In Handbook of Geometric Analysis, No. 1: Advanced Lectures in Mathematics, Vol. 7. International Press, 2008.

[Zel17] Steve Zelditch. Eigenfunctions of the Laplacian on a Riemannian manifold, volume 125 of CBMS Regional Conference Series in Mathematics. American Mathematical Soc., 2017.

[ZWX+18] Zhen Zhang, Mianzhi Wang, Yijian Xiang, Yan Huang, and Arye Nehorai. Retgk: Graph kernels based on return probabilities of random walks. Advances in Neural Information Processing Systems, 31, 2018.