Euler Characteristic Tools For Topological Data Analysis

Olympio Hacquard, Vadim Lebovici

Euler Characteristic Tools for Topological Data Analysis

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

Vadim Lebovici VADIM.LEBOVICI@INRIA.FR
Laboratoire de Mathématiques d'Orsay
Université Paris-Saclay, CNRS, Inria
91400 Orsay France

Editor:

Abstract

In this article, 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 so-called Euler characteristic profile. We show that this simple descriptor achieve state-of-the-art performance in supervised tasks at a very low 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. On the qualitative side, we provide numerous heuristics on the topological and geometric information captured by Euler profiles and their hybrid transforms. Finally, we prove stability results for these descriptors as well as asymptotic guarantees in random settings.

Keywords: Topological Data Analysis, Machine Learning, Multiparameter Persistence, Euler characteristic profiles, Hybrid transforms

1. Introduction

Extracting topological information from data of various natures follows a machinery that finds its origins in the works of Edelsbrunner et al. (2000). The main idea consists of building a one-parameter family of topological spaces on top of data and tracking the evolution of its topology, typically via homological computations. This multi-scale topological information is recorded in the form of what is called a *persistence diagram*; see Edelsbrunner et al. (2000); Edelsbrunner and Harer (2022). The space of persistence diagrams is a metric space for the so-called *bottleneck distance*, (Cohen-Steiner et al., 2007), but it cannot be isometrically embedded into a Hilbert space (Carrière and Bauer, 2018; Bubenik and Wagner, 2020). At the cost of losing some information, these diagrams are still often turned into vectors to perform various learning tasks such as classification, clustering, or regression. Most commonly used techniques include persistence images (Adams et al., 2017), landscapes (Bubenik et al., 2015), and more recently measure-oriented vectorizations in Royer et al. (2021) and neural network methods from Carrière et al. (2020); Reinauer et al. (2021). An overview of topological methods in machine learning has been presented in the survey of Hensel et al. (2021). These methods have demonstrated their efficiency in a wide variety of applications and types of data, such as health applications (Rieck et al., 2020; Fernández

and Mateos, 2022; Aukerman et al., 2021), biology (Ichinomiya et al., 2020; Rabadán and Blumberg, 2019) or material sciences (Lee et al., 2017; Hiraoka et al., 2016).

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 practice (Carlsson and Zomorodian, 2009). 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 Vipond (2020) and persistence images to the two-parameter setting in Carrière and Blumberg (2020). Besides their high level of sophistication, the main limitation of these tools is their computational cost; see Carrière and Blumberg (2020, Table 2) and Section 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 use them. First, they have demonstrated a good predictive power in various settings (Worsley et al., 1992; Richardson and Werman, 2014; Smith and Zavala, 2021; Jiang et al., 2020; Amézquita et al., 2022). Second, the simplicity of these descriptors translates into a reduced computational cost. They can be computed in linear time in the number of simplices in a simplicial filtration instead of typically matrix multiplication time for persistence diagrams (Milosavljević et al., 2011). Moreover, the locality of the Euler characteristic can be exploited to design highly efficient algorithms computing Euler curves, as in Heiss and Wagner (2017). 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 Adler and Taylor (2009), and asymptotic results of the Euler characteristic of a complex built on a Poisson process are established in Corollary 4.2 of Bobrowski and Adler (2014) and Corollary 6.2 of Bobrowski and Weinberger (2017). Furthermore, Euler curves associated with random point clouds are proven to be asymptotically normal for a well-chosen sampling regime in Krebs et al. (2021), where the authors also apply this construction to bootstrap. Fourth, they naturally generalise to the multi-parameter setting, becoming so-called *Euler characteristic surfaces* (Beltramo et al., 2022) and *profiles* (Dłotko and Gurnari, 2022).

We demonstrate that these tools reach state-of-the-art performance at a minimal computational cost when coupled with a powerful classifier such as a gradient boosting 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 Lebovici (2022) 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 (Lebovici, 2022, 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 (Lebovici, 2022, 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 Bobrowski and Kahle (2018) for a survey. Together with the works of Botnan and Hirsch (2022), our results form the first occurrence of limiting theorems in a multi-persistence framework in the literature.

Contributions and outline. In this article, we show that Euler characteristic profiles and their hybrid transforms are informative and highly efficient topological descriptors. Throughout the paper, we use classical methods based on persistence diagrams as a baseline for our descriptors. After introducing the necessary notions in Section 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 transforms can capture from data in Section 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 a gradient boosting (Sections 4.1, 4.2 and 4.4) at a very low computational cost (Section 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. They typically outperform persistence diagrams-based vectorizations, both in terms of accuracy and computational time.
- We demonstrate that hybrid transforms act as highly efficient information compressors and typically require a much smaller resolution than Euler profiles to reach a similar performance. They can also outperform Euler profiles in unsupervised classification tasks and in supervised tasks when plugging a linear classifier (Figure 7 and Sections 4.1 to 4.3). In Section 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 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. Similarly to persistence diagrams, we can 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 6).

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

2. Topological descriptors

This section presents all the necessary notions from simplicial geometry and the construction of the topological descriptors used throughout the article. 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 $f : \mathbb{R}^m \to \mathbb{C}$ 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 $\mathbb{R}^m$.

2.1 Simplicial complexes, 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 1 *Let $f : \mathcal{K} \to \mathbb{R}^m$ be a non-decreasing map for the inclusion of simplices, 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 popular example of simplicial complex is the Čech complex of a *point cloud*, that is, a finite subset of $\mathbb{R}^d$. This complex captures a lot of information on the geometry of the point cloud.

Example 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 such that 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* 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 alpha filtration*, which is a subcomplex of the Delaunay triangulation; see Bauer and Edelsbrunner (2017). See Figure 1 for an illustration.*

The properties of the Čech complex of a random point cloud have been deeply studied theoretically. We refer to Bobrowski and Kahle (2018) and Owada (2022) 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 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:*

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

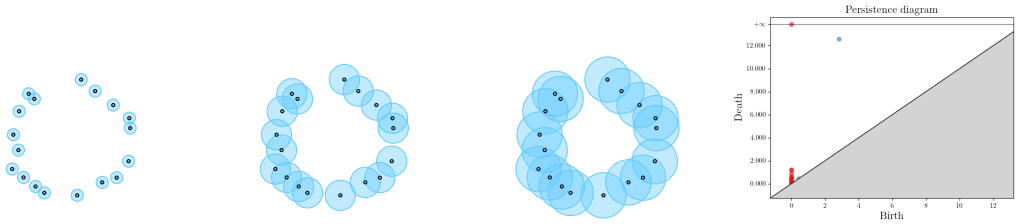
Again, we rather use function-alpha filtration *in numerical experiments, which are defined similarly using alpha complexes.*

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 2a 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, one denotes 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 (Lesnick and Wright, 2016). Note that sublevel-sets filtrations are one-critical. Conversely, any one-critical filtration is a sublevel set filtration for the function $f : \sigma \in \mathcal{K} \mapsto t(\sigma)$.

2.2 Persistence diagrams

Given a filtration of a simplicial complex, we want to extract multi-scale topological information from data. This is the objective of *persistent homology*, which constitutes the main tool of topological data analysis. It has found many practical applications (Rieck et al., 2020; Fernández and Mateos, 2022; Aukerman et al., 2021; Ichinomiya et al., 2020; Rabadán and Blumberg, 2019; Lee et al., 2017; Hiraoka et al., 2016) as well as applications to other fields of theoretical mathematics, such as symplectic geometry (Polterovich et al., 2020). This section introduces the basic objects of persistent homology as introduced in classical textbooks; see Edelsbrunner and Harer (2022); Oudot (2017). We try to keep the notions as intuitive as possible and do not lay out the technical details of homology theory.

The central tool of persistence theory is *homology*. Intuitively, given a topological space X, the k-th homology of X is a vector space whose dimension is equal to the number of independent k-dimensional holes of X. By 0-dimensional (resp. 1-dimensional, 2-dimensional) holes, we mean connected components (resp. cycles, voids). These k-dimensional holes are often called *homological features* of X. One can also define the homology of a simplicial



(a) $t = 0.4$ (b) $t = 1.5$ (c) $t = 2.1$ (d) Persistence diagram

Figure 1: Balls with varying radius $t > 0$ centered at each point of a finite subset $\mathbb{X} \subseteq \mathbb{R}^2$. These balls are used to define the Čech filtration $\check{\mathcal{C}}(\mathbb{X})$ and its corresponding persistence diagrams of dimension 0 (in red) and 1 (in blue).

complex $\mathcal{K}$, denoted by $H_k(\mathcal{K})$ for each integer $k \geq 0$, in such a way that they coincide with the above intuition when looking at the geometric realisation of the simplicial complex $\mathcal{K}$.

Given a one-parameter filtration $\mathcal{F}$ of a simplicial complex $\mathcal{K}$, one of the main properties of homology implies that for any $t \leq t'$ in $\mathbb{R}$, the inclusion of complexes $\mathcal{F}_t \subseteq \mathcal{F}_{t'}$ induces a linear map $H_k(\mathcal{F}_t) \to H_k(\mathcal{F}_{t'})$. The idea of *persistent homology* is to keep track of homological features appearing in the filtration through these maps. Each generator appears at some $a \in \mathbb{R}$ called its *birth* and disappears at some $b > a$ called its *death*. The couple $[a, b)$ is called the *bar* of the corresponding homological feature. The multiset of bars $[a, b)$ for each homological feature appearing in the filtration is called the *degree k persistence barcode* of $\mathcal{F}$. One can also represent this barcode as a multiset of points $(a, b) \in \mathbb{R}^2$ called the *degree k persistence diagram* of $\mathcal{F}$. We give an example of the construction of the persistence diagram of the Čech filtration in Figure 1d. In this case, the persistence diagram gives a lot of information on the topology and the geometry of the underlying point cloud. Here, when the radius of the balls is smaller than the smallest distance between any two points, we have as many connected components as points. As the radii of the balls grow, connected components of the union of balls merge (or die) one by one, except for one that never dies. Therefore, the degree 0 persistence diagram has only points born at 0. As for the degree 1 persistence diagram, a cycle appears when the radius of the balls is large enough and is filled approximately at the radius of the underlying circle, hence a single point in the persistence diagram.

2.3 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 (Beltramo et al., 2022; Dłotko and Gurnari, 2022) and topological and hybrid transforms of constructible functions (Schapira, 1995; Ghrist and Robinson, 2011; Lebovici, 2022). While these tools can be defined in the more general setting of o-minimal geometry, we focus on filtered simplicial complexes.

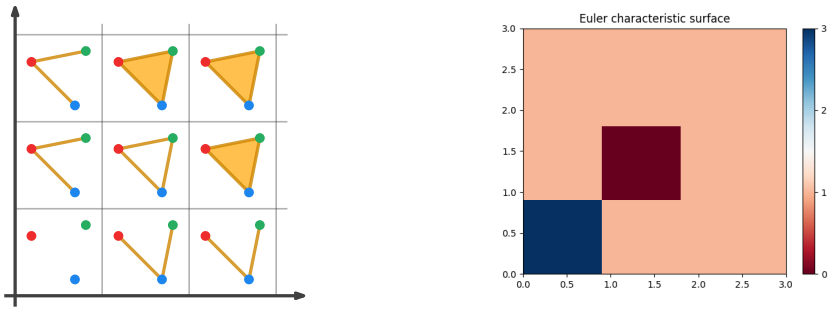
Given an m-parameter filtration, computing the Euler characteristic for every value of the parameter $t \in \mathbb{R}^m$ gives an integer-valued function on $\mathbb{R}^m$ that is a multi-scale descriptor of the evolution of the filtration with respect to t.

Definition 1 *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 Beltramo et al. (2022); Dłotko and Gurnari (2022).*

Figure 2 shows an Euler characteristic surface computed on an elementary example. Widely used in data analysis (Smith and Zavala, 2021; Dłotko and Gurnari, 2022; Beltramo et al., 2022; Jiang et al., 2020), this simple descriptor has proven to be efficient in capturing 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 2: 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 Schapira (1988-1989); Viro (1988):

Definition 2 *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-sets filtration, the pushforward operation has a simple expression:

Example 4 *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 $\xi_*\chi_f = \chi_{\xi \circ f}$.*

Hybrid transforms mixing Euler calculus and classical Lebesgue integration have been introduced in Lebovici (2022). These transforms are continuous and piecewise smooth and enjoy several beneficial properties, such as index theoretic formulae in the context of sublevel-sets 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 3 *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 4. 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 13).

Lemma 4 *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 and hybrid transforms constitute the two data descriptors we will use to perform topological data analysis. We give explicit expressions of these descriptors in specific cases below. These formulae will allow us to design algorithms to compute them in Section 3.1 and to build intuition on the type of behaviour they capture all along the paper.

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:

$$\chi_{\mathcal{F}} = \sum_{\sigma \in \mathcal{K}} (-1)^{\dim \sigma} \mathbf{1}_{Q_{t(\sigma)}}, \tag{2.1}$$

where $Q_u := \{t \in \mathbb{R}^m : t \geq u\}$ for any $u \in \mathbb{R}^m$.

Let $\kappa \in L^1(\mathbb{R})$. 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:

$$\psi_{\mathcal{F}}^{\kappa} : \xi \in \mathbb{R}_+^{m*} \longmapsto - \sum_{\sigma \in \mathcal{K}} (-1)^{\dim \sigma} \overline{\kappa} \left(\xi \cdot t(\sigma) \right). \tag{2.2}$$

Remark 5 *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 equal to the rescaled classical transform:

$$\psi_{\mathcal{F}}^{\kappa} : \xi \in \mathbb{R}_+^* \mapsto \xi \cdot \int_{\mathbb{R}} \kappa(\xi \cdot s) \chi_{\mathcal{F}}(s) \, \mathrm{d}s. \tag{2.3}$$

Assume now that $m \geq 2$. The hybrid transform with kernel κ differs from the classical integral transform:

$$\xi \in \mathbb{R}_+^{m*} \mapsto \int_{\mathbb{R}^m} \kappa(\xi \cdot x)\chi_{\mathcal{F}}(x)\,\mathrm{d}x.$$

We refer to Lebovici (2022, 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 (Lebovici, 2022, Examples 5.12 and 5.17). The interest in hybrid transforms over classical transforms can be motivated by the following example:

Example 5 *The one-parameter hybrid transform with kernel $\kappa(t) = \exp(-t)$ is also known as the* persistent magnitude *(Govc and Hepworth, 2021). It is used in O'Malley et al. (2023) as a new measure for estimating fractal dimensions of finite subsets $\mathbb{X} \subseteq \mathbb{R}^n$.*

2.4 Comparison of Euler characteristic tools with persistence diagrams

Suppose that $\mathcal{F}$ is a one-parameter filtration. In this case, Euler characteristic curves and hybrid transforms can simply be written as statistics of persistence diagrams. Denote the degree k persistence diagram of $\mathcal{F}$ by $\mathcal{D}_k = \{(a_i, b_i)\}_{i=1}^{n_k}$ where $-\infty < a_i^k < b_i^k \leq \infty$ and an integer $n_k \geq 0$. There exists k_0 such that persistence diagrams $\mathcal{D}_k$ are empty for all $k \geq k_0$. It is then straightforward to check that:

$$\chi_{\mathcal{F}} = \sum_{k\geq 0}\sum_{i=1}^{n_k}(-1)^k \mathbf{1}_{[a_i^k, b_i^k)}. \tag{2.4}$$

Let $\kappa \in L^1(\mathbb{R})$ and consider a primitive $\overline{\kappa}$ of κ. The hybrid transform with kernel κ of $\chi_{\mathcal{F}}$ therefore writes as:

$$\psi_{\mathcal{F}}^{\kappa} : \xi \in \mathbb{R}_+^* \mapsto \sum_{k\geq 0}\sum_{i=1}^{n_k}(-1)^k \left(\overline{\kappa}\left(\xi \cdot b_i^k\right) - \overline{\kappa}\left(\xi \cdot a_i^k\right)\right), \tag{2.5}$$

with the convention that $\overline{\kappa}(\xi \cdot b_i^k)$ is the limit of $\overline{\kappa}$ at $+\infty$ when $b_i^k = +\infty$. This connection between persistence diagrams and the one-parameter descriptors used in this article will be used for interpretation and in the asymptotic results of Section 6.

As we can see from (2.4) and (2.5), considering Euler curves and hybrid transforms instead of persistence diagrams implies a loss of information. More precisely:

- Both Euler characteristic curves and hybrid transforms can be written as an alternated sum over all homological degrees. As a consequence, the information contained in persistence diagrams is summed up across all homological degrees and reduced to a single descriptor.
- Even if only one persistence diagram is non-empty, the birth-death pairing of the points is lost while computing the Euler characteristic curve or hybrid transforms. In other words, Euler curves and hybrid transforms only depend on the sets $\{a_k^i\}$ and $\{b_k^i\}$ of all births and deaths respectively.
- Worse still, Euler curves are defined using indicator functions. As a consequence, the persistence diagrams $\{(0,1)\}$ and $\{(0,1/2),(1/2,1)\}$ share the same Euler curve and

the same hybrid transforms for all kernels. Therefore, the lifetime of a feature $b - a$, usually used as an indicator of the significance of the point (a, b) in the diagram is inaccessible.

The purpose of the following sections is to show that this loss of information does not result in a loss of accuracy when using Euler curves and hybrid transforms in machine learning tasks. Moreover, we show that the computation time is greatly reduced. This is due to the fact that Euler curves and hybrid transforms are computed using (2.1) and (2.2), bypassing the computation of homology and of persistence diagrams. This theoretical fact is backed up by experiments in Section 4.5. Finally, we use the fact that Euler curves and hybrid transforms can naturally be adapted to filtrations with more than one parameter, while there is no analogues of (2.4) and (2.5) in this case. We believe that these major gains indicate that such descriptors should be preferred over persistence diagrams when tackling machine learning problems.

3. Method

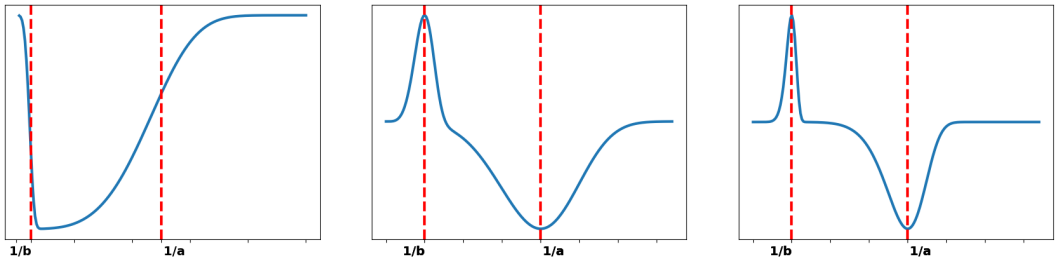
In this section, we describe the algorithms used to compute our descriptors and their implementation. We also give some intuition on choosing the kernel of hybrid transforms. Finally, we give heuristics on the type of information captured by Euler curves and their transforms on synthetic data sets.

3.1 Algorithm

In every experiment, and hence in our implementation, we restrict ourselves to one-critical filtrations. In that case, formulae (2.1) and (2.2) 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. 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 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 Milosavljević et al. (2011).

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 two different ways of choosing a sampling grid. The first method takes as input 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 method consists in giving as input a list $[(p_1, q_1), \ldots, (p_m, q_m)]$ with real numbers $0 \leq p_i < q_i \leq 1$. The algorithm then computes



(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 3: Hybrid transforms of $\chi_{\mathcal{F}} = \mathbf{1}_{[a,b)}$ for several choices of kernel κ

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 hybrid transforms, we provide 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. 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 $\overline{\kappa} : s \mapsto \exp(-x^p)$ and $\overline{\kappa} : s \mapsto x^p \exp(-x^p)$ by the paragraph *Kernel choice* below.

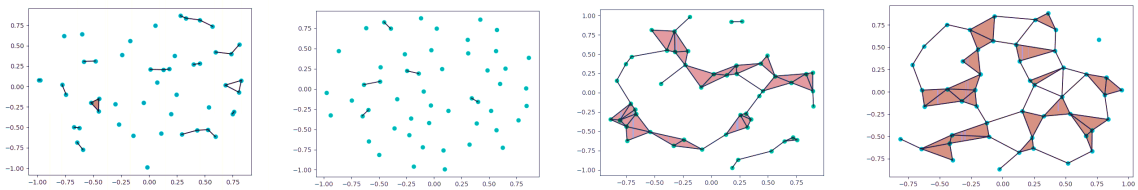
Kernel choice. To interpret integral transforms of Euler curves, we set $m = 1$ and compute them on the simple function $\chi_{\mathcal{F}} = \mathbf{1}_{[a,b)}$ with $a < b \in (0, +\infty)$. Recall that the hybrid transform has the simple expression (2.5). Figure 3 shows the hybrid transforms for several 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.

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 input data's topology, geometry, and sampling density. As already mentioned in Example 2, we instead use alpha filtration in numerical experiments for computational reasons.

Poisson and Ginibre point processes

While apparently coarse descriptors as opposed to persistence diagrams, Euler characteristic curves allow us to extract relevant scales at which topological differences between two different processes are revealed. To illustrate this claim, we try 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 Obayashi et al. (2018). Ginibre processes imply 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 Decreusefond and Moroz (2021). We classify this toy data set with a random forest classifier and select the two scales corresponding to maxima of the *feature importance* function of the classifier. In Figure 4, we plot two examples of point clouds together with their alpha complexes at these scales.



(a) PPP (b) GPP (c) PPP (d) GPP

Figure 4: 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 $t_1 < t_2$.

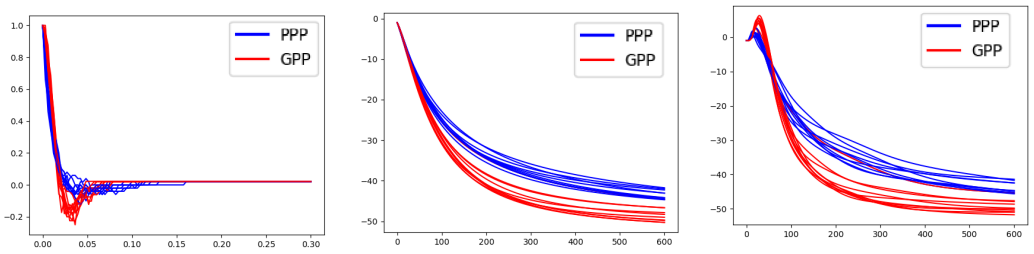
We plot Euler curves in Figure 5a. The Euler curves suggest that these classes differ at different scales, as it was visible in Figure 4:

- 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 allows for very large cycles to exist since there will typically be some large zones without any point, which is proscribed by GPP.

We remark that as opposed to persistence diagrams, our approach uses the birth times of edges instead of the usual degree 1 homological features. It seems that this information suffices to discriminate between the two classes.

We plot the transforms of these curves for several kernels in Figures 5b and 5c. 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 5: Euler characteristic curves and their transforms for PPP VS GPP data set

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, offering a finer analysis than persistence diagrams where all sampling-effects are represented as a jumbled clump of points near the diagonal. We consider two set-ups. The first one 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 (Diaconis et al., 2013). 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:

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

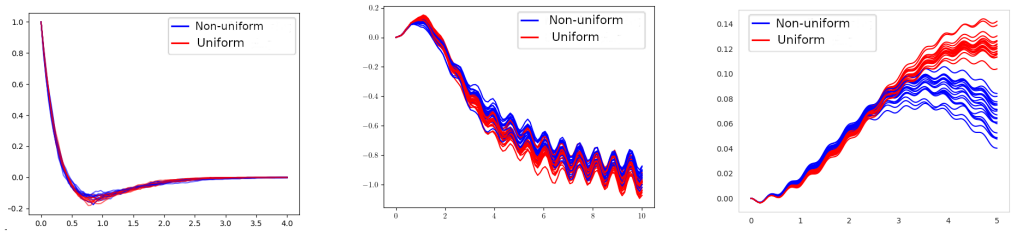
The second set-up consists of clouds of 500 points drawn in two ways on the unit sphere of $\mathbb{R}^3$. The first sampling is uniform. The second sampling is a non-uniform sampling where we draw θ uniformly on $[0, \pi]$ and φ according to a normal distribution centred on π. We obtain a point on the 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 6a and 6b, 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 6c, we show the hybrid transforms for the 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 sampling scheme shows up in the vertical shifts of the oscillations of the transforms. This interpretation allows us to go beyond the classical signal/noise dichotomy of persistence diagrams. Although it makes no doubt that this sampling information can be retrieved from the points close to the diagonal in the diagram, it is still unclear how to untangle the information on the sampling density itself and its support. 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 6: ECC and HT, two sampling on the torus and the sphere

Signal in clutter noise

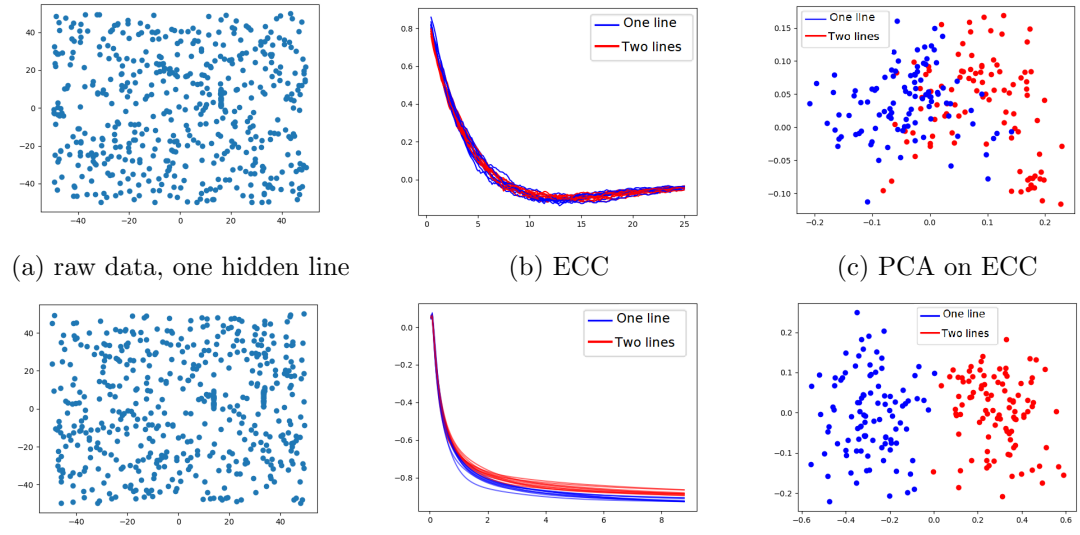
In this final illustrative experiment, we try to distinguish patterns in a heavy clutter noise. One class has one line hidden in the noise, while the other has two. Each line will induce a very dense zone creating early dying connected components. In Figure 7, we plot two examples of point clouds, the 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.

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`. Our timing experiments have been run on a workstation with an Intel(R) Core(TM) i7-4770 CPU (8 cores, 3.4GHz) and 8 GB RAM, running GNU/Linux (Ubuntu 20.04.1).

4.1 Curvature regression

We consider a set-up from Bubenik et al. (2020) 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



(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 7: Pattern hidden in clutter noise

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 the unit disk of the space with curvature $[-2, -1.96, \ldots, 1.96, 2]$ and validate our model on a testing set of 100 point clouds sampled from the disk of the space with random curvature drawn uniformly in $[-2, 2]$. We compare the R^2 scores in Table 1 with that of the original paper, which uses persistent landscapes (PL) along with a support vector regressor (SVR) and with Persformer (Reinauer et al., 2021). 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.

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 1: R^2 score for curvature regression data

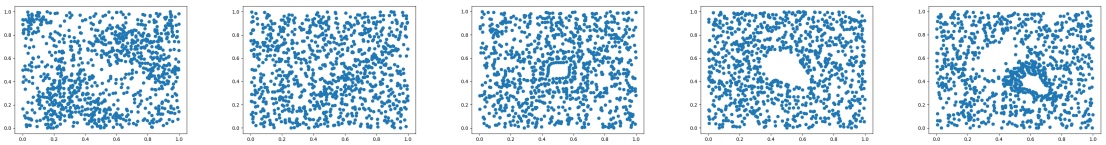
First, we remark that the ECC descriptor combined with a random forest has an accuracy comparable to state-of-the-art methods using persistence diagrams. We also remark that taking a transform does not improve the regression accuracy when considering a robust classifier such as RF but does improve the 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.

4.2 ORBIT5K data set

Supervised classification. The ORBIT5K data set is often used as a standard benchmark for classification methods in topological data analysis (Adams et al., 2017; Carrière et al., 2020; Reinauer et al., 2021). 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 generated recursively 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 8, 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) $\rho = 4.3$

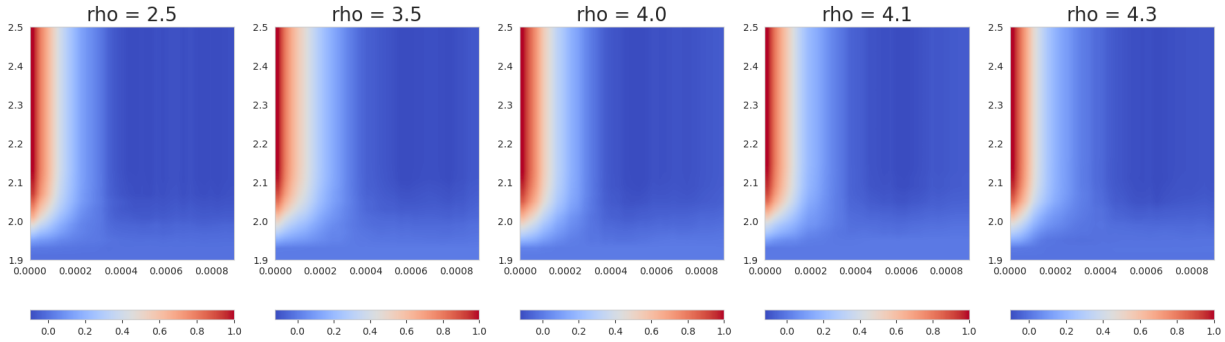
Figure 8: Examples of point clouds from the ORBIT5K data set.

Given an orbit of 1000 points, 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 accuracy scores with standard classification methods using persistence diagrams in Table 2. The results are averaged over ten runs. Sliced Wasserstein kernels (SW-K) and Persistence Fisher kernels (PF-K) are the two state-of-the-art kernel methods on persistence diagrams taken respectively from Carriere et al. (2017) and Le and Yamada (2018). Perslay and Persformer are two methods that use a neural network architecture to vectorize persistence diagrams (Carrière et al., 2020; Reinauer et al., 2021). The Euler characteristic curves and one-parameter hybrid transforms (HT1) are computed on the alpha filtration of the point cloud. The Euler characteristic surfaces and two-dimensional 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 classified using the XGBoost classifier (Chen and Guestrin, 2016). We select the hyperparameters of our descriptors by cross-validation:

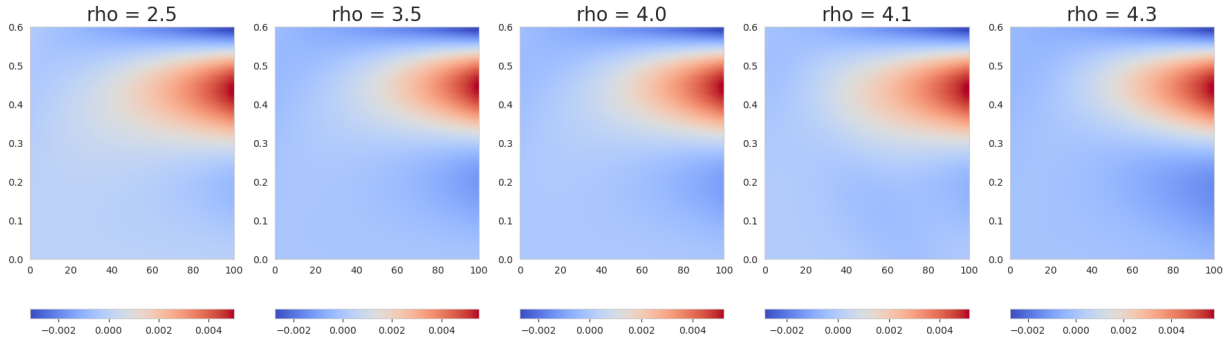
- For the ECC, the quantiles (see *Implementation* in Section 3.1) are selected in $\{(0.1, 0.9), (0.2, 0.8), (0.3, 0.7)\}$.
- For the ECS, the quantiles are selected in the same set as for the ECC for both parameters.
- For the HT1, the range is selected in $\{[0, 50], [0, 100], [0, 500], [0, 1000]\}$ 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 $\{[0, 50], [0, 80], [0, 100], [0, 500]\}$.

We show in Figure 9 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 $\overline{\kappa} : s \mapsto s^4 \exp(-s^4)$.



(a) ECS



(b) HT2

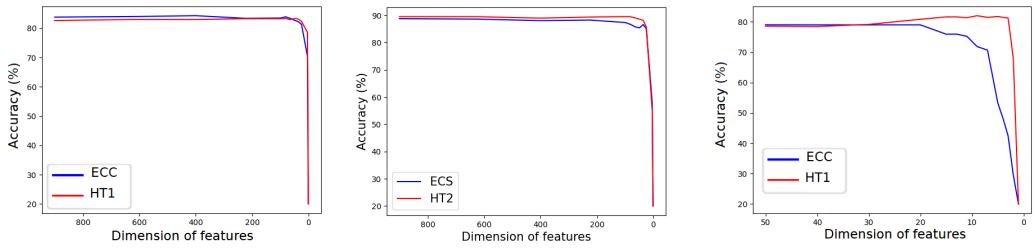
Figure 9: Examples of 2D descriptors

Method	SW-K	PF-K	Perslay	Persformer
Accuracy	83.6 ± 0.9	85.9 ± 0.8	87.7 ± 1.0	91.2 ± 0.8
Method	ECC + XGB	HT1 + XGB	**ECS + XGB**	HT2 + XGB
Accuracy	83.8 ± 0.5	82.8 ± 1.4	**91.8 ± 0.4**	89.9 ± 0.5

Table 2: Classification scores for the `ORBIT5K` data set

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 4.5.

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 10. 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

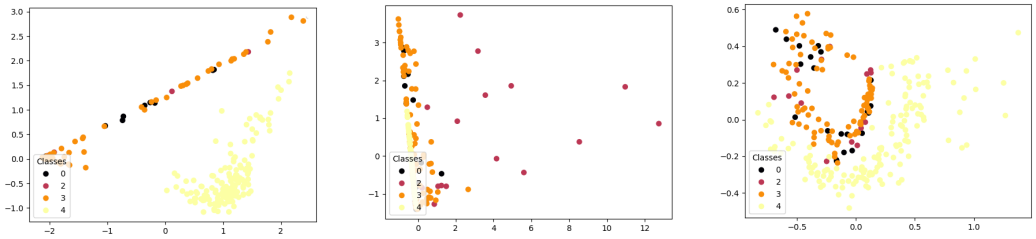
Figure 10: Accuracy with respect to feature dimension.

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 (De Deuge et al., 2013) traditionally used for multi-class classification. Likewise to Section 4.2, all descriptors are computed using a function-alpha filtration associated with a kernel density estimator post-composed with a decreasing function.

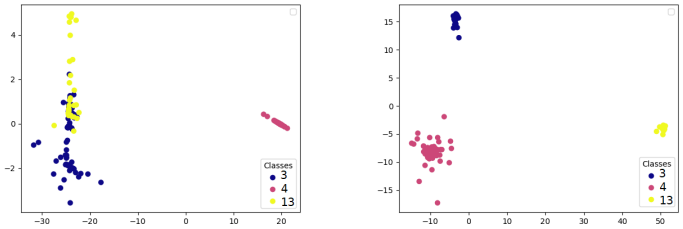
Unsupervised setting. In Figure 11, we show a PCA of the ECSs and HTs on the classes *4-wheeler vehicles* (labelled 0), *buses* (2), *cars* (3), and *pedestrians* (4). In this case, the ECSs separate the class of pedestrians from all the vehicle classes. The same separation is achieved by the HTs with 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 12. 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 by 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 11: PCA plots of ECSs and HTs for the Sidney object recognition data set.



(a) ECS (b) HTs, $\overline{\kappa}(s) = s^4 \exp(-s^4)$

Figure 12: LDA plots of ECSs and HTs for the Sidney object recognition data set.

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 introduced in Sun et al. (2009) and defined as follows. For a graph $\mathcal{G} = (V, E)$, the *HKS function with diffusion parameter* t is defined for each $v \in V$ by:

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

where λ_k is the k-th eigenvalue of the normalized graph Laplacian and ψ_k the corresponding eigenfunction. We consider the HKS with parameters $t = 1$ and $t = 10$ as filtrations. We also consider the $1/2-$Ricci and Forman curvatures (Samal et al., 2018), 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) 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 $\{s \mapsto \cos(s), s \mapsto \cos(s^2), s \mapsto \exp(-s^4), s \mapsto s^4 \exp(-s^4)\}$ for HTn. We report our scores in Table 3. The first four methods are state-of-the-art classification methods on graphs that use kernels or neural networks. We report

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)
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 3: Mean accuracy and standard deviation for graph data.

the scores from the original papers, Tran et al. (2019); Zhang et al. (2018); Verma and Zhang (2017); Xu et al. (2019). Perslay (Carrière et al., 2020), and Atol (Royer et al., 2021) 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 3.1 and 4.5. Second, as opposed to other invariants such as multi-parameter persistent images (Carrière and Blumberg, 2020), 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.

4.5 Timing

In this section, we compare the computational cost of our different methods to that of persistence images, a well-known vectorization of persistence diagrams introduced in Adams et al. (2017) and generalized to the multi-parameter setting in Carrière and Blumberg (2020). 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; see (Carrière and Blumberg, 2020, Table 2).

Constant resolution. We report in Table 4 the time to compute our descriptors and persistent images on the full `ORBIT5K` data set with a fixed resolution of 900. We assume that simplex trees are precomputed[1] using the `Gudhi` library (Rouvreau, 2015). Our descriptors are computed using the parameters achieving the highest accuracy for the classification

1. 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 codensity on point clouds.

task; see Section 4.2. Persistence images are computed with the `Gudhi` library for one-parameter filtrations and with the `MMA` package for two-parameter filtrations (Loiseaux et al., 2022b) 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 of simplex trees in the one-parameter case or persistence approximations in the two-parameter case (Loiseaux et al., 2022a, 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	HT2	PI2
16	719	139	144	805	2034

Table 4: Computation times (s) for `ORBIT5K` with constant resolution.

As expected from the time complexities of the algorithms (Section 3.1), Euler characteristic profiles are at least ten times faster than persistence images to compute, and hybrid transforms are four times faster in the two-parameter case. One-parameter hybrid transforms may appear costly to compute, but this point 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 5 the time to compute our descriptors on the full `ORBIT5K` data set with the lowest resolution before accuracy drop-out as reported in Figure 10. 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 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	HT2
16	5	135	69

Table 5: Computation times (s) for `ORBIT5K` with smallest resolution before accuracy drop-out.

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. Hybrid transforms have

similar accuracy but are more costly to compute, especially in the one-parameter setting; see Table 4. 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 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 profiles (Figure 10) at a meagre computational cost (Table 5). This makes hybrid transforms robust descriptors combining dimension reduction and feature extraction.

Theoretically, multi-parameter hybrid transforms benefit from their expression as one-parameter ones (Lemma 4). This allows us to prove almost sure convergence results under some mild assumptions in Section 6.

4.7 Extensions

We have validated our method on simplicial complexes built on point clouds and graph data. Nonetheless, the methodology described in this paper 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 Smith and Zavala (2021); Jiang et al. (2020). 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 Richardson and Werman (2014); Beltramo et al. (2022); Dłotko and Gurnari (2022), a thorough benchmark against other persistence methods and state-of-the-art image processing methods is still missing. Moreover, 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 Xian et al. (2022).

5. Stability properties

The success of topological data analysis inherits from the stability theorem for persistence diagrams from Cohen-Steiner et al. (2007). Loosely speaking, it means that under mild assumptions, small changes in the filtration function imply small changes in the diagram. Such results are crucial to designing consistent estimators; see, for instance, Bobrowski et al. (2017). Over the past decade, more distances on persistence diagrams have been introduced. Inspired by optimal transport theory, the notion of p-Wasserstein distance is introduced by Cohen-Steiner et al. (2010) where a stability result is also proven. A finer stability result for the p-Wasserstein distance can be found in Skraba and Turner (2020). In addition, several stability results for Euler characteristic tools have been derived in Curry et al. (2022); Dłotko and Gurnari (2022); Perez (2022).

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 so-called *signed barcodes*. As a corollary, we bound the L^q norms of hybrid transforms by the same quantity. To continue our comparison with persistence diagrams, we prove that in the one-parameter case, the signed 1-Wasserstein distance between signed barcodes is bounded from above by the well-known 1-Wasserstein distance between persistence diagrams.

The notions of signed barcodes and of signed 1-Wasserstein distance have been introduced in Oudot and Scoccola (2021) and are recalled below. We follow the same conventions as in Oudot and Scoccola (2021, Section 2) for the definitions of multisets and bijections between them. The rest of the section is devoted to the statement of our stability results. All proofs are written in Section 7.1.

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 finitely generated filtrations (Lemma 14). We denote by $\mathrm{FP}(\mathbb{R}^m)$ the group of finitely presented functions over $\mathbb{R}^m$. These functions have a kind of diagram (or barcode) that can be used to define an analogue of the 1-Wasserstein distance. A *decomposition* of $\varphi \in \mathrm{FP}(\mathbb{R}^m)$ is a couple $(\mathcal{B}^+, \mathcal{B}^-)$ of finite multisets of points in $\mathbb{R}^m$ such that:

$$\varphi = \sum_{u \in \mathcal{B}^+} \mathbf{1}_{Q_u} - \sum_{v \in \mathcal{B}^-} \mathbf{1}_{Q_v}.$$

Such a decomposition always exists, and there is a unique $\overline{\mathcal{B}} = (\mathcal{B}^+, \mathcal{B}^-)$ such that $\mathcal{B}^+ \cap \mathcal{B}^- = \varnothing$, called the *signed barcode of* φ; see (Oudot and Scoccola, 2021, Proposition 13). While two different notions of signed barcode are defined in loc. cit., we focus here on the so-called *minimal Hilbert decomposition signed barcode*.

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 $(\mathcal{B}^+, \mathcal{B}^-)$ and $(\mathcal{B}'^+, \mathcal{B}'^-)$, the *signed 1-Wasserstein distance* between them is:

$$\widehat{d}_1(\varphi, \varphi') = \inf \left\{ \varepsilon > 0 : \exists \text{ bijection } h : \mathcal{B}^+ \cup \mathcal{B}'^- \to \mathcal{B}^- \cup \mathcal{B}'^+ \text{ with } \mathrm{cost}(h) \leq \varepsilon \right\}.$$

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 prove stability results involving functional norms on Euler characteristic profiles and their hybrid transforms. The case $m = 1$ is well known for 1-Wasserstein distance on persistence diagrams; see Curry et al. (2022, Lemma 4.10), Dłotko and Gurnari (2022, Proposition 3.2).

Proposition 6 *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 $m = 1$:

$$\|\chi_{\mathcal{F}} - \chi_{\mathcal{F}'}\|_1 \leq \widehat{d}_1(\chi_{\mathcal{F}}, \chi_{\mathcal{F}'})\,.$$

This stability result for Euler characteristic profiles implies a similar stability result for hybrid transforms, as stated in the following corollary.

Corollary 7 *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:*

$$\|\psi^\kappa_{\mathcal{F}} - \psi^\kappa_{\mathcal{F}'}\|_{L^q_K} \leq C_{K,q}\, \|\kappa\|_\infty\, \widehat{d}_1(\chi_{\mathcal{F}}, \chi_{\mathcal{F}'})\,.$$

Now, we prove two connections of the signed 1-Wasserstein distance with more classical distances between filtrations. The first connection is made with the 1-Wasserstein between persistence diagrams (Cohen-Steiner et al., 2010). We start by recalling it. Denote by $\mathcal{D}$ and $\mathcal{D}'$ the degree k persistence diagrams of $\mathcal{F}$ and $\mathcal{F}'$. The *p-Wasserstein distance* between $\mathcal{D}$ and $\mathcal{D}'$ is defined as:

$$W_p(\mathcal{D}, \mathcal{D}') = \inf_\eta \left(\sum_{x \in \mathcal{D}} \|x - \eta(x)\|^p \right)^{1/p}$$

where the infimum is taken over all bijections $\eta : \mathcal{D} \cup \Delta \to \mathcal{D}' \cup \Delta$ where $\Delta = \{(s, s) | s \in \mathbb{R}\}$ is the diagonal of $\mathbb{R}^2$. This definition allows for matchings between diagrams with different number of points. We can now state the following connection between the 1-Wasserstein distance on diagrams and the signed 1-Wasserstein on Euler characteristic curves.

Lemma 8 *Let $\mathcal{F}$ and $\mathcal{F}'$ be two finitely generated one-parameter filtrations of respective simplicial complexes $\mathcal{K}$ and $\mathcal{K}'$. Denote their respective persistence diagrams $\mathcal{D}$ and $\mathcal{D}'$. Then,*

$$\widehat{d}_1(\chi_{\mathcal{F}}, \chi_{\mathcal{F}'}) \;\leq\; 2 \sum_{k \geq 0} W_1\big(\mathcal{D}, \mathcal{D}'\big).$$

Combined with Proposition 6 and Corollary 7, this lemma ensures that L^1 norms of Euler characteristic curves and L^q norms of their hybrid transforms are controlled by a classical distance between their persistence diagrams. This is another element of comparison between Euler characteristic tools and persistence diagrams. It is important to note that all homology degrees have to be taken into account for the result to hold.

The second connection is established between the signed 1-Wasserstein distance on Euler characteristic profiles and L^1 norms on filtration functions defined on the same simplicial complex, as stated by the lemma below. It has already been formulated in a slightly different form in Dłotko and Gurnari (2022, 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 9 *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 functions is an indicator of their sensitivity to the underlying geometry. Persistent images (Adams et al., 2017) share this property, while persistence landscapes (Bubenik et al., 2015; Vipond, 2020) do not, as they are controlled by the L^∞ distance between functions.

6. Statistical properties

This section provides statistical guarantees for our descriptors computed on a random sample, as the sample size tends to infinity.

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 10 is a pointwise law of large numbers, while Theorem 12 establishes 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 demonstrate that as the sampling density appears explicitly in Theorems 10 and 12, hybrid transforms can, at least asymptotically, be used to discriminate between samples from different probability densities.

Theorem 10 *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 $A_0, \ldots, A_{d-1}$ on $\mathbb{R}_+^*$ that depend only on $\overline{\kappa}$ such that for every $\xi > a$,*

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

We defer the proof to Section 7.2. Note that a law of large numbers for the Euler characteristic curve has been established in Corollary 6.2 of Bobrowski and Weinberger (2017) for all possible regimes and could be integrated to derive a similar result for hybrid transforms. However, this result is only established for the Čech filtration over a uniform sampling on the flat torus. As we want to show the dependency over the sampling density, we have adapted the stronger limit results of Owada (2022) for persistence diagrams. It is therefore a key assumption that we are in the so-called *sparse regime*, that is, $nr_n^d \to 0$. To make this

law of large numbers more understandable, 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: the Poisson regime $n^{k+2} r_n^{d(k+1)} \to c > 0$ and the vanishing regime $n^{k+2} r_n^{d(k+1)} \to 0$.

Theorem 10 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 $k = 0, 1, \ldots, d-1$ 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 Krebs et al. (2021), which we recall hereafter for the sake of completeness. First, recall that a function h on $\mathbb{R}^m$ is *blocked* if it can be written $h = \sum_{i=1}^{m^d} b_i \mathbf{1}_{A_i}$ where $b_1, \ldots, b_{m^d}$ are non-negative real numbers and the A_i are axis-aligned rectangles in $\mathbb{R}^m$. 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 11 (Krebs et al., 2021, Theorem 3.4) *Let $T > 0$ and $X_1, \ldots, X_n$ 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 Krebs et al. (2021) 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 12 *Consider the setting of Theorem 11. 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:*

$$\sqrt{n}\left(\psi^\kappa_{\mathcal{F}_n} - \mathbb{E}\left[\psi^\kappa_{\mathcal{F}_n}\right]\right) \underset{n\to\infty}{\longrightarrow} \tilde{\mathfrak{G}} \quad a.s.,$$

in $\left(\mathcal{C}^0[a,M], \|\cdot\|_\infty\right)$. Furthermore, the covariance of the limiting process is defined by:

$$\mathbb{E}\left[\tilde{\mathfrak{G}}(\xi_1)\tilde{\mathfrak{G}}(\xi_2)\right] = \xi_1\xi_2 \int_0^{T/\xi_1}\int_0^{T/\xi_2} \kappa(\xi_1 t)\,\kappa(\xi_2 s)\,\mathrm{cov}\left(\mathfrak{G}(s), \mathfrak{G}(t)\right) \mathrm{d}s\,\mathrm{d}t,$$

where $\mathfrak{G}$ is the Gaussian process defined in Theorem 11.

6.2 Limit theorem for multi-parameter hybrid transforms

Here, we adopt the sampling model of Hiraoka et al. (2018). Consider a point process Φ on $\mathbb{R}^d$ and its restriction Φ_L to $[-L/2, L/2]^d$. Let $\mathcal{S}(\mathbb{R}^d)$ be the collection of all finite (non-empty) subsets in $\mathbb{R}^d$, to be thought of as the set of all simplices. Let $f = (f_1, \ldots, f_m) : \mathcal{S}(\mathbb{R}^d) \to [0, \infty]^m$ be a measurable function, non-decreasing with respect to the inclusions of faces. According to Example 1, f induces a filtration on every simplicial complex of $\mathbb{R}^d$. The following theorem derives a law of large numbers for the hybrid transform in the multi-parameter case.

Theorem 13 *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 $(x, y) \in (\mathbb{R}^d)^2$,*

$$\|x - y\| \leq \rho\left(f_i(\{x, y\})\right). \tag{6.1}$$

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 Hiraoka et al. (2018) 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 4 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 Penrose and Yukich (2001) 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 Botnan and Hirsch (2022). Alpha and function-Čech filtrations that we used in our experiments do not verify this assumption.

As pointed out in Hiraoka et al. (2018, Example 1.3), Čech and Vietoris-Rips filtrations satisfy (6.1) for $\rho : t \mapsto 2t$. We provide below two examples of families a broad family of multi-parameter filtrations satisfying (6.1).

Example 6 *It is easy to check that the function-alpha filtration considered in the applications of Sections 4.2 and 4.3 satifies* (6.1).

We give another class of filtrations satisfying (6.1) that contains in particular the distance-to-measure (DTM) filtrations (Anai et al., 2020).

Example 7 *Let h be a positive and bounded function from $\mathbb{R}^d$ to $\mathbb{R}$. The weighted Čech complex introduced in Anai et al. (2020) 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* (6.1) *for $\rho : t \mapsto \max(\max h, 2t)$.*

7. Proofs

In this section, we prove the results stated in Sections 5 and 6.

7.1 Proofs of stability results

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:

$$\widehat{d}_1(\varphi, \varphi') = \inf\left\{\varepsilon > 0 : \exists \text{ bijection } h : \mathcal{C}^+ \cup \mathcal{C}'^- \to \mathcal{C}^- \cup \mathcal{C}'^+ \text{ with } \mathrm{cost}(h) \leq \varepsilon\right\}. \tag{7.1}$$

7.1.1 Profiles of finitely generated filtrations are finitely presented

The following lemma is well-known. We prove it for completeness. Recall that the k-th *Betti function* of a finitely generated filtration $\mathcal{F}$ is defined as the function $\beta_{\mathcal{F},k} : t \in \mathbb{R}^m \mapsto \dim H_k(\mathcal{F}_t)$.

Lemma 14 *Let $\mathcal{F}$ be a finitely generated m-parameter filtration. The k-th Betti function $\beta_{\mathcal{F},k}$ is finitely presented and $\chi_{\mathcal{F}} = \sum_{k \in \mathbb{Z}} (-1)^k \beta_{\mathcal{F},k}$. In particular, the Euler characteristic profile of $\mathcal{F}$ is finitely presented.*

Proof The fact that $\beta_{\mathcal{F},k}$ is finitely presented follows from the fact that the family of vector spaces $(H_k(\mathcal{F}_t))_{t \in \mathbb{R}^m}$ forms a *finitely presented m-parameter persistence module* (see Lemma A). This last fact is well known but goes beyond the scope of the paper and is not explicitly written elsewhere in the literature. We provide a proof in Appendix A. The equality between the alternated sum of Betti functions of $\mathcal{F}$ and its Euler characteristic profile follows from the classical formula for the Euler characteristic of any simplicial complex $\mathcal{K}$ stating that:

$$\chi(\mathcal{K}) = \sum_{k \geq 0} (-1)^k \dim H_k(\mathcal{K}).$$

The fact that $\chi_{\mathcal{F}}$ is finitely presented is then straightforward. ■

The signed 1-Wasserstein distance between Euler characteristic profiles is bounded from above by the same distance between Betti functions, as stated in the lemma below. It will be crucial to proving the other results.

Lemma 15 *Let $\mathcal{F}$ and $\mathcal{F}'$ be two finitely generated m-parameter filtrations of simplicial complexes $\mathcal{K}$ and $\mathcal{K}'$ respectively. Then,*

$$\widehat{d}_1(\chi_{\mathcal{F}}, \chi_{\mathcal{F}'}) \leq \sum_{k\geq 0} \widehat{d}_1\big(\beta_{\mathcal{F},k}, \beta_{\mathcal{F}',k}\big)\,.$$

Proof By Lemma 14, the functions $\beta_{\mathcal{F},k}$ and $\beta_{\mathcal{F}',k}$ are all finitely presented. A collection of decompositions $(\mathcal{B}_k^+, \mathcal{B}_k^-)$ of $\beta_{\mathcal{F},k}$ for all $k \in \mathbb{N}$ induces a decomposition $(\mathcal{B}^+, \mathcal{B}^-) = (\cup_k \mathcal{B}_k^+, \cup_k \mathcal{B}_k^-)$ of $\chi_{\mathcal{F}}$. A similar decomposition $(\mathcal{B}'^+, \mathcal{B}'^-)$ of $\chi_{\mathcal{F}'}$ is induced by decompositions $(\mathcal{B}_k'^+, \mathcal{B}_k'^-)$ of $\beta_{\mathcal{F}',k}$ for all $k \in \mathbb{N}$. Moreover, a collection of bijections of multisets $h_k : \mathcal{B}_k^+ \cup \mathcal{B}_k'^- \to \mathcal{B}_k^- \cup \mathcal{B}_k'^+$ for all $k \in \mathbb{N}$ induces a bijection of multisets $h : \mathcal{B}^+ \cup \mathcal{B}'^- \to \mathcal{B}^- \cup \mathcal{B}'^+$ with $\mathrm{cost}(h) = \sum_{k\in\mathbb{N}} \mathrm{cost}(h_k)$. Taking the infimum over all bijections h_k yields the result by (7.1). ■

7.1.2 Proof of Lemma 8

The degree k persistence diagram of $\mathcal{F}$ is given by $\mathcal{D}_k = \big\{\big(a_i^k, b_i^k\big)\big\}_{i=1,\dots,n_k}$ for real numbers $-\infty < a_i^k < b_i^k \leq \infty$ and an integer $n_k \geq 0$. This diagram induces a decomposition $(\mathcal{A}_k, \mathcal{B}_k) = (\{a_i^k\}_i, \{b_i^k\}_i)$ of $\beta_{\mathcal{F},k}$. Similarly, the degree k persistence diagram $\mathcal{D}_k' = \Big\{\big(a_j'^k, b_j'^k\big)\Big\}_{j=1,\dots,n_k'}$ of $\mathcal{F}'$ induces a decompositon $(\mathcal{A}_k', \mathcal{B}_k') = (\{a_j'^k\}_j, \{b_j'^k\}_j)$ of $\beta_{\mathcal{F}',k}$. Moreover, a partial matching M between $\mathcal{D}_k$ and $\mathcal{D}_k'$ induces a bijection of multisets $h : \mathcal{A}_k' \cup \mathcal{B}_k \to \mathcal{A}_k \cup \mathcal{B}_k'$ defined by $h(a') = a$ and $h(b) = b'$ when $((a,b),(a',b')) \in M$, by $h(b) = a$ when (a,b) is unmatched and by $h(a') = b'$ when (a',b') is unmatched. Moreover, the cost of the matching M and the cost of the bijection h satisfy $\mathrm{cost}(h) \leq 2\,\mathrm{cost}(M)$. Taking the infimum over all partial matching M, one has $\widehat{d}_1\big(\beta_{\mathcal{F},k}, \beta_{\mathcal{F}',k}\big) \leq W_1(H_k\mathcal{F}, H_k\mathcal{F}')$. Lemma 15 yields the result.

7.1.3 Proof of Proposition 6

Recall that $m \geq 1$. Consider decompositions $(\mathcal{B}^+, \mathcal{B}^-)$ and $(\mathcal{B}'^+, \mathcal{B}'^-)$ of $\chi_{\mathcal{F}}$ and $\chi_{\mathcal{F}'}$ respectively. Assume there is a bijection $h : \mathcal{B}^+ \cup \mathcal{B}'^- \to \mathcal{B}^- \cup \mathcal{B}'^+$. If no such bijection exists, then $\widehat{d}_1(\chi_{\mathcal{F}}, \chi_{\mathcal{F}'})$ is infinite, and the inequality trivially holds. One has:

$$\chi_{\mathcal{F}} - \chi_{\mathcal{F}'} = \sum_{u\in\mathcal{B}^+\cup\mathcal{B}'^-} \mathbf{1}_{Q_u} - \sum_{v\in\mathcal{B}^-\cup\mathcal{B}'^+} \mathbf{1}_{Q_v} = \sum_{u\in\mathcal{B}^+\cup\mathcal{B}'^-} \mathbf{1}_{Q_u} - \mathbf{1}_{Q_{h(u)}}.$$

Therefore,

$$\|\chi_{\mathcal{F}} - \chi_{\mathcal{F}'}\|_{1,M} \leq \sum_{u\in\mathcal{B}^+\cup\mathcal{B}'^-} \|\mathbf{1}_{Q_u} - \mathbf{1}_{Q_{h(u)}}\|_1. \tag{7.2}$$

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 (7.2) and the fact that $\|\mathbf{1}_{[u,v)}\|_1 = |u - v|$.

7.1.4 Proof of Corollary 7

It follows from the definition of hybrid transforms that:

$$\|\psi^{\kappa}_{\mathcal{F}} - \psi^{\kappa}_{\mathcal{F}'}\|_{L^q_K} \leq \begin{cases} \|\kappa\|_\infty \displaystyle\int_K \int_{\mathbb{R}} |\xi_*\chi_{\mathcal{F}}(s) - \xi_*\chi_{\mathcal{F}'}(s)| \,\mathrm{d}s \,\mathrm{d}\xi & \text{if } q \in [1,\infty), \\ \|\kappa\|_\infty \sup\limits_{\xi\in K} \displaystyle\int_{\mathbb{R}} |\xi_*\chi_{\mathcal{F}}(s) - \xi_*\chi_{\mathcal{F}'}(s)| \,\mathrm{d}s & \text{if } q = \infty. \end{cases}$$

Moreover, Proposition 6 with $m = 1$ ensures that for any $\xi \in K$,

$$\|\xi_*\chi_{\mathcal{F}} - \xi_*\chi_{\mathcal{F}'}\|_1 \leq \widehat{d}_1(\xi_*\chi_{\mathcal{F}}, \xi_*\chi_{\mathcal{F}'})\,.$$

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 $(\mathcal{B}^+, \mathcal{B}^-)$ and $(\mathcal{B}'^+, \mathcal{B}'^-)$ of $\chi_{\mathcal{F}}$ and $\chi_{\mathcal{F}'}$ respectively. They induce decompositions $(\xi_*\mathcal{B}^+, \xi_*\mathcal{B}^-)$ and $(\xi_*\mathcal{B}'^+, \xi_*\mathcal{B}'^-)$ of $\xi_*\chi_{\mathcal{F}} = \chi_{\xi_*\mathcal{F}}$ and $\xi_*\chi_{\mathcal{F}'} = \chi_{\xi_*\mathcal{F}'}$ respectively by the formula $\xi_*\mathcal{B}^\pm = \{\xi \cdot u : u \in \mathcal{B}^\pm\}$ and a similar one for $\mathcal{F}'$. Consider a bijection of multisets $h : \mathcal{B}^+ \cup \mathcal{B}'^- \to \mathcal{B}^- \cup \mathcal{B}'^+$. It induces a bijection of multisets $\xi_*h : \xi_*\mathcal{B}^+ \cup \xi_*\mathcal{B}'^- \to \xi_*\mathcal{B}^- \cup \xi_*\mathcal{B}'^+$ defined by $\xi \cdot u \mapsto \xi \cdot h(u)$ with cost:

$$\mathrm{cost}(\xi_*h) = \sum_{t\in\xi_*\mathcal{B}^+\cup\xi_*\mathcal{B}'^-} \|t - \xi_*h(t)\|_1 = \sum_{u\in\mathcal{B}^+\cup\mathcal{B}'^-} \|\xi\cdot u - \xi\cdot h(u)\|_1 \leq \|\xi\|_\infty \cdot \mathrm{cost}(h).$$

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 (7.1).

7.1.5 Proof of Lemma 9

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 cost $\mathrm{cost}(h) = \sum_{\sigma\in\mathcal{K}} \|f(\sigma) - g(\sigma)\|_1 = \|g - f\|_1$. The result follows from (7.1).

7.2 Proofs of statistical results

In this section, we prove the asymptotic results for the hybrid transforms stated in Section 6.

7.2.1 Proof of Theorem 10

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 $n \to \infty$.

Let us define $\Delta := \{(x,y) : 0 \leq x \leq y < \infty\} \cup \{(x,\infty) : 0 \leq x < \infty\}$ and for every (s,t,u,v) such that $0 \leq s \leq t \leq u \leq v \leq \infty$, denote by $R_{s,t,u,v}$ the rectangle $(s,t] \times (u,v]$ of Δ. Recall that a finite persistence diagram $\mathcal{D} = \cup_{i=1}^l (a_i, b_i)$ 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 Owada (2022) 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:

$$\frac{1}{n^{k+2} r_n^{d(k+1)}} \mu_{k,n} \xrightarrow[n\to\infty]{v} \frac{1}{(k+2)!} \left(\int_{\mathbb{R}^d} g^{k+2}(x)\, \mathrm{d}x \right) \mu_k \quad \text{a.s.}, \tag{7.3}$$

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 (Owada, 2022, Sec. 3.1), which does not depend on g and such that the measure μ_k is defined by:

$$\mu_k(R_{s,t,u,v}) = \int_{\mathbb{R}^{d(k+1)}} H_{s,t,u,v}(0, y_1, \ldots, y_{k+1})\, \mathrm{d}y_1 \ldots \mathrm{d}y_{k+1}.$$

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 (2.5), 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 (7.3) 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 $A_k(\xi) = \int_\Delta h_\xi \mathrm{d}\mu_k$.

7.2.2 Proof of Theorem 12

Let $T > 0$ such that κ is supported in $[0, T]$. Let $a, M > 0$ and let $\xi \in [a, M]$. According to (2.3), 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 11, 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:

$$\sqrt{n} \left(\chi_{\mathcal{F}_n}(t) - \mathbb{E}[\chi_{\mathcal{F}_n}(t)] \right) \xrightarrow[n\to\infty]{} \mathfrak{G}(t), \tag{7.4}$$

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 (7.4) 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 $\mathfrak{G}$.

7.2.3 Proof of Theorem 13

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:

$$\rho'\left(\sum_{j=1}^m \xi_j f_j(\{x,y\})\right) \geq \rho'(\xi_i f_i(\{x,y\}) \geq \rho(f_i(\{x,y\})) \geq \|x-y\|. \tag{7.5}$$

The filtration function $\xi \cdot f$ therefore verifies all the hypotheses of Theorem 1.5 of Hiraoka et al. (2018), 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 (7.5) to hold. As in the proof of Theorem 10, 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 (2.5) together with Lemma 4, we have that:

$$\psi^\kappa_{\mathcal{F}_L}(\lambda\xi) = \sum_{k=0}^{d-1}(-1)^k \langle \mu_{k,L}^{\xi_*\mathcal{F}}, h_\lambda \rangle.$$

Hence the result, by the vague convergence $\frac{1}{L^d}\mu_{k,L}^{\xi_*\mathcal{F}} \xrightarrow{v} \nu_k^{\xi\cdot f}$ for every $k \in [\![0, d-1]\!]$.

Acknowledgments

The authors are grateful to Steve Oudot, François Petit, Clément Levrard, Wolfgang Polonik and Mathieu Carrière for useful discussions. The authors are also grateful to the Inria Datashape team for comments on an early version of this work at the annual seminar and to the anonymous reviewer for their insightful comments that helped greatly improve the exposition.

Appendix A

In this appendix, we prove that a finitely generated filtration has finitely presentable persistent homology. As explained in Section 7.1, this fact is well-known. Its proof is included for completeness. We follow the same notations and conventions as in Oudot and Scoccola (2021, Section 2).

Lemma A *Let $\mathcal{F}$ be a finitely generated m-parameter filtration of a simplicial complex $\mathcal{K}$ and let $k \geq 0$. The m-parameter persistence module $H_k(\mathcal{F})$ is finitely presentable. In particular, its Hilbert function is finitely presented.*

Proof Since $\mathcal{F}$ is finitely generated, the support of any $\sigma \in \mathcal{K}$ has a finite number of minimal elements. The set of these elements is called the *births* of σ and denoted by $\mathrm{birth}(\sigma)$. Since $\mathcal{K}$ is finite and $\mathcal{F}$ is finitely generated, there is a finite subset $G = I_1 \times \ldots \times I_m \subseteq \mathbb{R}^m$ such that $\mathrm{birth}(\sigma) \subseteq G$ for any $\sigma \in \mathcal{K}$.

Given a persistence module M over $\mathbb{R}^m$, we denote by $r(M)$ its restriction to G. Given a persistence module N over G, the *extension of N* is the persistence module $e(N)$ over $\mathbb{R}^m$ defined by:

$$e(N)(t) = N\left(\max\{g \in G : g \leq t\}\right).$$

This defines functors r and e between the category of persistence modules over $\mathbb{R}^m$ to the category of persistence modules over G and conversely. It is an easy exercise to check that these functors are exact.

We prove that $H_k(\mathcal{F}) \simeq e \circ r(H_k(\mathcal{F}))$. It is well known—see Bauer and Scoccola (2022, Lemma 5) for a proof—that this implies that $H_k(\mathcal{F})$ is finitely presentable. Recall that *barcodes* of free persistence modules are defined in Oudot and Scoccola (2021, Section 2) and denote by $\mathcal{C}_i(\mathcal{F})$ the free persistence module with barcode $\bigcup_\sigma \mathrm{birth}(\sigma)$ where the union is taken over all simplices $\sigma \in \mathcal{K}$ of dimension i. Consider the diagram:

$$\mathcal{C}_{k+1}(\mathcal{F}) \xrightarrow{\partial_k} \mathcal{C}_k(\mathcal{F}) \xrightarrow{\partial_{k-1}} \mathcal{C}_{k-1}(\mathcal{F}),$$

where the maps ∂_k and ∂_{k-1} are induced by the boundary operator from simplicial homology. The persistence module $H_k(\mathcal{F})$ is then the homology of the above diagram, i.e.,

$$H_k(\mathcal{F}) = \mathrm{Im}(\partial_k)/\,\mathrm{Ker}(\partial_{k-1}).$$

For any $i \in \mathbb{N}$, the definition of G and of $\mathcal{C}_i(\mathcal{F})$ implies that $\mathcal{C}_i(\mathcal{F}) \simeq e \circ r(\mathcal{C}_i(\mathcal{F}))$. The result then follows from the fact that $e \circ r$ is an exact functor and hence commutes with computing homology.

We are left to prove that the Hilbert function of $H_k(\mathcal{F})$ is finitely presented. Since the persistence module $H_k(\mathcal{F})$ is finitely presentable, it admits a finite free resolution:

$$0 \to F_m \to \cdots \to F_0 \to H_k(\mathcal{F}) \to 0. \tag{A.1}$$

See for instance Botnan and Lesnick (2022, Section 7.2) for more details. Each free module with barcode $\mathcal{B}(F_i)$ has a finitely presented Hilbert function:

$$\mathrm{Hil}(F_i) = \sum_{t \in \mathcal{B}(F_i)} \mathbf{1}_{Q_t}.$$

Now, exactness of the sequence (A.1) ensures that:

$$\mathrm{Hil}(H_k(\mathcal{F})) = \sum_{i=0}^{m} (-1)^i \mathrm{Hil}(F_i),$$

hence the result. ■

References

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.

Robert J. Adler and Jonathan E. Taylor. *Random fields and geometry*. Springer Science & Business Media, 2009.

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.

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.

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.

Ulrich Bauer and Herbert Edelsbrunner. The morse theory of čech and delaunay complexes. *Transactions of the American Mathematical Society*, 369(5):3741–3762, 2017.

Ulrich Bauer and Luis Scoccola. Generic two-parameter persistence modules are nearly indecomposable. arXiv preprint:2211.15306, 2022.

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.

Omer Bobrowski and Robert J. Adler. Distance functions, critical points, and the topology of random cech complexes. *Homology, Homotopy and Applications*, 16(2):311–344, 2014.

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

Omer Bobrowski and Shmuel Weinberger. On the vanishing of homology in random čech complexes. *Random Structures & Algorithms*, 51(1):14–51, 2017.

Omer Bobrowski, Sayan Mukherjee, and Jonathan E. Taylor. Topological consistency via kernel estimation. *Bernoulli*, 23(1):288 – 328, 2017. doi: 10.3150/15-BEJ744. URL `https://doi.org/10.3150/15-BEJ744`.

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.

Magnus Bakke Botnan and Michael Lesnick. An introduction to multiparameter persistence. arXiv preprint:2203.14289. To appear in: Proceedings of the ICRA 2020, 2022.

Peter Bubenik and Alexander Wagner. Embeddings of persistence diagrams into hilbert spaces. *Journal of Applied and Computational Topology*, 4(3):339–351, 2020.

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

Peter Bubenik et al. Statistical topological data analysis using persistence landscapes. *Jounral of Machine Learning Research*, 16(1):77–102, 2015.

Gunnar Carlsson and Afra Zomorodian. The Theory of Multidimensional Persistence. *Discrete & Computational Geometry*, 42(1):71–93, 2009. ISSN 0179-5376.

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.

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

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

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

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.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37:103–120, 2007.

David Cohen-Steiner, Herbert Edelsbrunner, John Harer, and Yuriy Mileyko. Lipschitz functions have l p-stable persistence. *Foundations of computational mathematics*, 10(2): 127–139, 2010.

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.

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.

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.

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.

Paweł Dłotko and Davide Gurnari. Euler characteristic curves and profiles: a stable shape invariant for big data problems. arXiv preprint:2212.01666, 2022.

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

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.

Ximena Fernández and Diego Mateos. Topological biomarkers for real-time detection of epileptic seizures. *arXiv preprint:2211.02523*, 2022.

Robert Ghrist and Michael Robinson. Euler–Bessel and Euler–Fourier transforms. *Inverse Problems*, 27(12), 2011.

Dejan Govc and Richard Hepworth. Persistent magnitude. *Journal of Pure and Applied Algebra*, 225(3), 2021.

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.

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

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.

Yasuaki Hiraoka, Tomoyuki Shirai, and Khanh Duy Trinh. Limit theorems for persistence diagrams. *The Annals of Applied Probability*, 28(5):2740–2780, 2018.

Takashi Ichinomiya, Ippei Obayashi, and Yasuaki Hiraoka. Protein-folding analysis using features obtained by persistent homology. *Biophysical Journal*, 118(12):2926–2937, 2020.

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.

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.

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

Vadim Lebovici. Hybrid transforms of constructible functions. *Foundations of Computational Mathematics*, pages 1–47, 2022.

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.

Michael Lesnick and Matthew Wright. Interactive visualization of 2-d persistence modules. 2016.

David Loiseaux, Mathieu Carriere, and Andrew J Blumberg. Efficient approximation of multiparameter persistence modules. *arXiv preprint:2206.02026*, 2022a.

David Loiseaux, Mathieu Carrière, and Hannah Schreiber. Multipersistence Modules Approximation (`MMA`). `https://github.com/DavidLapous/multipers`, 2022b.

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.

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.

Miguel O'Malley, Sara Kalisnik, and Nina Otter. Alpha magnitude. *Journal of Pure and Applied Algebra*, 227(11):107396, 2023.

Steve Oudot and Luis Scoccola. On the stability of multigraded betti numbers and hilbert functions. arXiv preprint:2112.11901, 2021.

Steve Y Oudot. *Persistence theory: from quiver representations to data analysis*, volume 209. American Mathematical Soc., 2017.

Takashi Owada. Convergence of persistence diagram in the sparse regime. *The Annals of Applied Probability*, 32(6):4706–4736, 2022.

Mathew D Penrose and Joseph E Yukich. Central limit theorems for some graphs in computational geometry. *Annals of Applied probability*, pages 1005–1041, 2001.

Daniel Perez. Euler and betti curves are stable under wasserstein deformations of distributions of stochastic processes. 2022.

Leonid Polterovich, Daniel Rosen, Karina Samvelyan, and Jun Zhang. *Topological persistence in geometry and analysis*, volume 74. American Mathematical Society, 2020.

Raúl Rabadán and Andrew J Blumberg. *Topological data analysis for genomics and evolution: topology in biology*. Cambridge University Press, 2019.

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

Eitan Richardson and Michael Werman. Efficient classification using the euler characteristic. *Pattern Recognition Letters*, 49:99–106, 2014.

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.

Vincent Rouvreau. Alpha complex. In *GUDHI User and Reference Manual*. GUDHI Editorial Board, 2015. URL `http://gudhi.gforge.inria.fr/doc/latest/group__alpha__complex.html`.

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.

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.

Pierre Schapira. Cycles lagrangiens, fonctions constructibles et applications. *Séminaire Équations aux dérivées partielles (Polytechnique) dit aussi "Séminaire Goulaouic-Schwartz"*, 1988-1989.

Pierre Schapira. Tomography of constructible functions. In *International Symposium on Applied Algebra, Algebraic Algorithms, and Error-Correcting Codes*, pages 427–435. Springer, 1995.

Primoz Skraba and Katharine Turner. Wasserstein stability for persistence diagrams. *arXiv preprint arXiv:2006.16824*, 2020.

Alexander Smith and Victor M Zavala. The euler characteristic: A general topological descriptor for complex data. *Computers & Chemical Engineering*, 154:107463, 2021.

Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, volume 28, pages 1383–1392. Wiley Online Library, 2009.

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.

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.

Oliver Vipond. Multiparameter persistence landscapes. *Journal of Machine Learning Research*, 21(61):1–38, 2020.

Oleg Yanovich Viro. Some integral calculus based on euler characteristic. In *Topology and geometry—Rohlin seminar*, pages 127–138. Springer, 1988.

Keith J Worsley, Alan C Evans, Sean Marrett, and P Neelin. A three-dimensional statistical analysis for cbf activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism*, 12(6):900–918, 1992.

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.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

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.