Estimation and Quantization of Expected Persistence Diagrams

Vincent Divol [*1] **Théo Lacombe** [*1]

Abstract

Persistence diagrams (PDs) are the most common descriptors used to encode the topology of structured data appearing in challenging learning tasks; think e.g. of graphs, time series or point clouds sampled close to a manifold. Given random objects and the corresponding distribution of PDs, one may want to build a statistical summary—such as a mean—of these random PDs, which is however not a trivial task as the natural geometry of the space of PDs is not linear. In this article, we study two such summaries, the Expected Persistence Diagram (EPD), and its quantization. The EPD is a measure supported on $\mathbb{R}^2$, which may be approximated by its empirical counterpart. We prove that this estimator is optimal from a minimax standpoint on a large class of models with a parametric rate of convergence. The empirical EPD is simple and efficient to compute, but possibly has a very large support, hindering its use in practice. To overcome this issue, we propose an algorithm to compute a quantization of the empirical EPD, a measure with small support which is shown to approximate with near-optimal rates a quantization of the theoretical EPD.

1. Introduction

Topological data analysis (TDA) is a modern field in data science which has found a variety of succesful domains of application such as material science (Saadatfar et al., 2017; Buchet et al., 2018), cellular data (Cámara, 2017), social graph classification (Zhao & Wang, 2019; Carrière et al., 2020), shape analysis (Li et al., 2014; Carrière et al., 2015) to name a few. It provides a machinery to encode the topological properties (such as the presence of connected components, loops, cavities, etc.) of a structured object in a multi-scale fashion. Relying on persistent homology theory (Edelsbrunner et al., 2000; Zomorodian & Carlsson, 2005; Edelsbrunner & Harer, 2010), its main output is a descriptor called a *persistence diagram* (PD): it is a discrete measure $\sum_{i\in I} \delta_{x_i}$ (roughly, a set of points) supported on the open half-plane $\Omega = \{(t_1, t_2) \in \mathbb{R}^2,\ t_2 > t_1\}$, where each point x_i of the PD accounts in a quantitative way for the presence of a topological feature in a given object. The space of PDs, $\mathcal{D}$, is equipped with an *optimal partial transport* metric OT_p, where $1 \leq p \leq \infty$, which shares similarities with the so-called Wasserstein metric W_p used in the optimal transport literature (Villani, 2008; Santambrogio, 2015).

Statistics with PDs. In applications, one is generally led to consider a sample of several PDs, say $\mu_1, \dots, \mu_n$, encoding the topology of some underlying phenomenon generating the different observations. Assuming that these PDs are sampled i.i.d. according to some underlying distribution P, it is natural to search for some characteristic quantities to describe P. As the space of PDs $(\mathcal{D}, \mathrm{OT}_p)$ is not a vector space, but only a metric space, even building elementary statistics is a difficult task. For instance, approximating Fréchet means (a.k.a. barycenters) of a sample of PDs with respect to OT_p metrics requires to develop specific techniques (Turner et al., 2014; Lacombe et al., 2018; Vidal et al., 2019), while their exact computation is intractable. An alternative is to embed the PDs in a Hilbert or Banach space, using explicit vectorizations (Bubenik, 2015; Adams et al., 2017) or implicit through kernel methods (Reininghaus et al., 2015; Carrière et al., 2017), then using standard statistical and learning tools. However, such embeddings do not preserve the metric structure of the space of PDs (Bubenik & Wagner, 2020; Wagner, 2021) nor the interpretability of PDs. In comparison, the *expected persistence diagram* (EPD) $\mathbf{E}(P)$ of a distribution P of PDs lies in a natural metric extension of the space of PDs while its empirical counterpart can be computed faithfully. Originally introduced in (Divol & Chazal, 2019), the EPD is a measure on Ω which associates to each set $A \subset \Omega$ the expected number of points which belongs to A in the random diagrams $\mu \sim P$. The properties of this object were studied in (Divol & Chazal, 2019; Divol & Lacombe, 2020).

Contributions. We consider the situation where one has access to a n-sample of PDs $\mu_1, \dots, \mu_n$ following some (unknown) law P. A natural way to estimate the EPD of P

[*]Equal contribution [1]Université Paris-Saclay, CNRS, Inria, Laboratoire de Mathématiques d'Orsay, 91405, Orsay, France. Correspondence to: Vincent Divol <vincent.divol@inria.fr>, Théo Lacombe <theo.lacombe@inria.fr>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

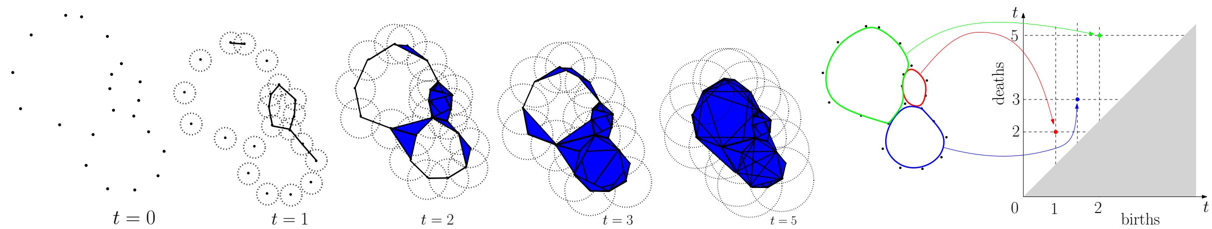

Figure 1. Čech filtration on a 2D point cloud in dimension $D=1$ (recording loops) and the corresponding PD.

is to consider its empirical counterpart, which simply reads $\overline{\mu}_n := \frac{1}{n}(\mu_1 + \cdots + \mu_n)$. By leveraging techniques from optimal transport theory, we show in Section 3 that $\overline{\mu}_n$ approximates $\mathbf{E}(P)$ at the parametric rate $n^{-1/2}$ with respect to the loss OT_p^p under non-restrictive assumptions, and that it is optimal from a minimax perspective. In practice, the support of the measure $\overline{\mu}_n$ is obtained as the union of the support of each diagram and tends to be very large if $n \gg 1$, hindering the use of this empirical descriptor in applications. To overcome this issue, we propose in Section 4 an online algorithm to compute a quantization of the empirical EPD and show that—provided a good initialization—the output of our algorithm approximates a quantization of the EPD at an appropriate rate. For the sake of conciseness, proofs have been deferred to the supplementary material along with code to reproduce our experiments.

Related Work. Divol & Chazal (2019) show that under mild assumptions the EPD is a measure with density supported on the half-plane Ω, and propose an estimation procedure of the EPD based on kernel density estimation. However, they defined convergence in terms of L_2 metrics between densities instead of the more natural diagram metric OT_p considered in this work and did not exhibit rates of convergence. In optimal transport literature, the study of convergence rates between a measure and its empirical counterpart for the Wasserstein distance W_p dates back to (Dudley, 1969), while more recent papers (Singh & Póczos, 2018; Fournier & Guillin, 2015; Kloeckner, 2020; Lei et al., 2020) provide tight controls of the convergence rate of the quantity W_p^p. There are however two main differences between this line of results and our framework. First, despite both being optimal transport metrics, there exist key differences between the metric OT_p and the Wasserstein metric W_p (see Section 2). Furthermore, we are not in the common situation where one observes i.i.d. realizations $X_1, \dots, X_n$ in Ω and considers the empirical measure $\frac{1}{n}(\delta_{X_1} + \cdots + \delta_{X_n})$ but in the more general setting where one observes measures $\mu_1, \dots, \mu_n$ on Ω following some law P and considers the distance between the expected measure $\mathbf{E}(P)$ and its empirical counterpart $\frac{1}{n}(\mu_1 + \cdots + \mu_n)$.

The problem of quantization of measures, namely approximating a given measure with another measure with support of fixed size, has been studied in depth when those measures are supported on $\mathbb{R}^d$ equipped with its natural Euclidean geometry, see for instance (Graf & Luschgy, 2007; Fischer, 2010; Levrard et al., 2015; Bourne et al., 2018). In the context of PDs, where the quantization problem is generally referred to as computing *codebooks* or *bag-of-words* (Zieliński et al., 2018; 2020), existing methods propose to quantize PDs running a k-mean algorithm on the diagram points. The intuition that points in a diagram that are close to the boundary $\partial\Omega$ of the half-plane Ω represent less important topological features is taken into account through the introduction of weight functions, requiring to introduce an important hyper-parameter whose choice is unclear in general. Our approach differs from the latter on two aspects: first, we do not quantize a single diagram (should it be a superposition of diagrams as in (Zieliński et al., 2020)) but work in an *online* fashion with a sequence of observed diagrams. Second, we work with the standard diagram metric OT_p. In doing so, we directly take the boundary $\partial\Omega$ into account in the formulation of our problem without needing to introduce a weight function. Our quantization algorithm significantly builds on (Chazal et al., 2021, Alg. 2). The main difference is that Chazal et al. intend to quantize a measure with respect to the 2-Wasserstein distance on $\mathbb{R}^d$, while we work with the metric OT_p on $\Omega \subset \mathbb{R}^2$. This change of perspective introduces some specificities in our problem and allows us to derive results more suited to the context of persistence diagrams. Furthermore, while standard algorithms work with $p=2$, we propose a simple variation to encompass the case $p=+\infty$, central in TDA as one retrieves the so-called *bottleneck* distance.

2. Background

Persistence diagrams (PDs). Let X be a topological space and let $f : X \to \mathbb{R}$ be a real-valued continuous function. The *sublevel sets* of (X, f) are defined as $\mathcal{F}_t := \{w \in X,\ f(w) < t\}$. As the scale parameter t increases from $-\infty$ to $+\infty$, one observes a nested sequence of sets called the *filtration* of X by f. Given a fixed dimension D, persistent homology (see (Edelsbrunner & Harer, 2010) for an introduction) provides tools to record the scales at which a topological feature (a connected component for

$D = 0$, a loop for $D = 1$, a cavity for $D = 2$, etc.) appears or disappears in the sublevel sets. For instance, a loop (one-dimensional topological component) might appear at some scale t_1 (its birth time) in the sublevel set $\mathcal{F}_{t_1}$, and disappear ("get filled") at some scale $t_2 > t_1$. One says that the loop *persists* over the interval $[t_1, t_2]$. This results in a collection of intervals[1]—each of them accounting for the presence of a topological feature recorded in the filtration process—that can be encoded as a multiset supported on the open half-plane $\Omega = \{x = (t_1, t_2),\ t_2 > t_1\} \subset \mathbb{R}^2$, or, equivalently, as a locally finite discrete measure $\mathrm{Dgm}(f) := \sum_i \delta_{x_i}$, where δ_{x_i} denotes the Dirac mass located at $x_i \in \Omega$. Of particular interest is the case where $X = \mathbb{R}^d$, and $f : w \in \mathbb{R}^d \mapsto \mathrm{dist}(w, A)$ is the distance function to A a compact subset of $\mathbb{R}^d$ (for instance a point cloud), see Figure 1. The corresponding diagram, called the Čech persistence diagram of A, will be denoted by $\mathrm{Dgm}(A)$.

Metrics for PDs. Let $\|\cdot\|$ be the Euclidean norm and let $\mathrm{spt}(\mu)$ denote the support of a measure μ. Let $\partial\Omega := \{(t, t),\ t \in \mathbb{R}\}$ be the diagonal (which is also the boundary of Ω), and $\overline{\Omega} := \Omega \sqcup \partial\Omega$. Given $1 \leq p < \infty$, and two measures μ, ν supported on Ω, one can define the distance between μ and ν using an *optimal partial transport* metric:

$$\mathrm{OT}_p(\mu, \nu) := \inf_{\pi \in \mathrm{Adm}(\mu,\nu)} \left(\iint_{\overline{\Omega} \times \overline{\Omega}} \|x - y\|^p \mathrm{d}\pi \right)^{\frac{1}{p}}, \quad (2.1)$$

where $\mathrm{Adm}(\mu, \nu)$ is the set of measures supported on $\overline{\Omega} \times \overline{\Omega}$ whose first (resp. second) marginal coincides with μ (resp. ν) on Ω (note in particular that π is not constrained on $\partial\Omega \times \partial\Omega$). The definition is extended to $p = \infty$ by replacing $\left(\iint_{\overline{\Omega} \times \overline{\Omega}} \|x - y\|^p \mathrm{d}\pi \right)^{\frac{1}{p}}$ by $\sup\{\|x - y\|,\ (x, y) \in \mathrm{spt}(\pi)\}$, and the distance OT_∞ is called the *bottleneck distance*, central in TDA due to its strong stability properties (Cohen-Steiner et al., 2007; Chazal et al., 2016). Let $\|x - \partial\Omega\| = (t_2 - t_1)/\sqrt{2}$ be the *persistence* of a point $x = (t_1, t_2) \in \Omega$, that is its distance to the diagonal $\partial\Omega$. The space $(\mathcal{M}^p, \mathrm{OT}_p)$ of *persistence measures* is defined as the space of (non-negative) Radon measures μ supported on Ω that have finite *total persistence*, i.e. $\mathrm{Pers}_p(\mu) := \int \|x - \partial\Omega\|^p \mathrm{d}\mu(x) < \infty$ (this condition ensures that OT_p is always finite). Note that the distance OT_p is not only defined for PDs (elements of $\mathcal{D}$), but for measures on Ω with arbitrary support, therefore making it possible to define a similarity notion between a PD and a more general measure such as an EPD, a crucial aspect of this work.

The metrics OT_p are similar to the *Wasserstein distances* used in optimal transport (Santambrogio, 2015, Ch. 5): for σ, τ two measures *having the same total mass* on a metric space (S, ρ), the distance $W_{p,\rho}(\sigma, \tau)$ is defined as the infimum of $\left(\int_{S^2} \rho(x, y)^p \mathrm{d}\pi(x, y) \right)^{1/p}$ over all transport plans π between σ and τ, i.e. measures on $S \times S$ which have for first (resp. second) marginal σ (resp. τ). When ρ is the Euclidean distance we write W_p instead of $W_{p,\rho}$. Despite those similarities, there is however a crucial difference between the Wasserstein distance and the OT_p distance: the constraints in (2.1) only involves the marginals on Ω, allowing us to transport mass to and from the boundary of the space $\partial\Omega$. It makes, in particular, the distance OT_p between measures of different total masses well-defined. The metrics OT_p were introduced by Figalli & Gigli (2010) as a way to study the heat equation with Dirichlet boundary conditions, but Divol & Lacombe (2020) observed that these metrics actually coincide with the standard metrics used to compare persistence diagrams (Edelsbrunner & Harer, 2010, Ch. 8).

Expected persistence diagrams. Let P be a probability distribution supported on $(\mathcal{M}^p, \mathrm{OT}_p)$. Let $\mathbf{E}(P)$ be the measure defined by, for $A \subset \Omega$ compact,

$$\mathbf{E}(P)(A) := \mathbb{E}_P[\mu(A)], \quad (2.2)$$

where $\mu \sim P$, and $\mu(A)$ is the (random) number of points of μ that belongs to A. This deterministic measure, called the *expected persistence diagram* (EPD) of P, was introduced in (Divol & Chazal, 2019) were authors proved that, under mild assumptions, it admits a density with respect to the Lebesgue measure on Ω. Importantly, the EPD is a persistence measure but not a PD in general.

3. Minimax Estimation of the EPD

Let P be a distribution of PDs, and $\mathbf{E}(P)$ be its EPD. Given a n-sample $\mu_1, \ldots, \mu_n$ of law P, the empirical EPD is defined as $\overline{\mu}_n := \frac{1}{n} \sum_i \mu_i$. In this section, we control the distance $\mathrm{OT}_p^p(\overline{\mu}_n, \mathbf{E}(P))$ under moment assumptions on the underlying law P. Note that, according to (Divol & Lacombe, 2020, Thm. 3.7) and the law of large numbers, $\overline{\mu}_n \xrightarrow{\mathrm{OT}_p} \mathbf{E}(P)$ almost surely under the minimal assumption that $\mathbb{E}_P[\mathrm{Pers}_p(\mu)] < \infty$ (see Lemma 3 in the supplementary material). Our goal here is to understand the rate at which this convergence holds.

Let A_L be the ℓ_1-ball in $\mathbb{R}^2$ centered at $(-L/\sqrt{8}, L/\sqrt{8})$ of radius $L/\sqrt{2}$. For $0 \leq q \leq \infty$ and $L, M > 0$, we let $\mathcal{M}^q_{L,M}$ be the set of measures $\mu \in \mathcal{M}^q$ which are supported on A_L, with $\mathrm{Pers}_q(\mu) \leq M$. Let $\mathcal{P}^q_{L,M}$ be the set of probability distributions which are supported on $\mathcal{M}^q_{L,M}$. It is known that persistence diagrams belong to the set $\mathcal{M}^q_{L,M}$ under non-restrictive assumptions. Namely, we have the following result.

Lemma 1 (Cohen-Steiner et al. (2010))**.** *Let X be a d-dimensional compact Riemannian manifold, and let $f : X \to \mathbb{R}$ be a Lipschitz continuous function. Then, for*

[1] In the greatest generality, there may be some intervals of the form $[t_1, +\infty)$. In the following, such intervals are simply discarded if ever present.

every $q > d$, $\mathrm{Dgm}(f)$ belongs to $\mathcal{M}^q_{L,M}$ for some L, M depending on X, q and the Lipschitz constant of f.

In particular, for $q > 0$, no constraints on the total number of points of the persistence diagram are imposed. This is particularly interesting in applications, where the number of points in PDs is likely to be large, while their total persistence Pers_q may be moderate, see e.g. (Divol & Polonik, 2019) for asymptotics in the case of the Čech persistence diagrams of large samples on the cube.

Theorem 1. *Let $1 \leq p < \infty$ and $0 \leq q < p$. Let $P \in \mathcal{P}^q_{L,M}$ and let $\mu_1, \ldots, \mu_n$ be a n-sample from law P. If $\overline{\mu}_n$ is the associated empirical EPD, then,*

$$\mathbb{E}[\mathrm{OT}_p^p(\overline{\mu}_n, \mathbf{E}(P))] \leq cML^{p-q}\left(\frac{1}{n^{1/2}} + \frac{a_p(n)}{n^{p-q}}\right), \quad (3.1)$$

where c depends on p and q, and $a_p(n) = 1$ if $p > 1$, $\log(n)$ if $p = 1$.

In particular, if $p \geq q + 1/2$, we obtain a parametric rate of convergence of $n^{-1/2}$. This is always the case if $q = 0$, i.e. if we assume that all the diagrams sampled according to P have less than M points. According to Lemma 1, it is also the case if $\mu_i = \mathrm{Dgm}(f_i)$ for some random 1-Lipschitz functions $f_i : X \to \mathbb{R}$, where X is a d-dimensional compact Riemannian manifold with $p > d + 1/2$.

From a statistical perspective, it is natural to wonder if better estimates of $\mathbf{E}(P)$ exist. A possible way to answer this question is given by the minimax framework. Let $\mathcal{P}$ be a set of probability distributions on $\mathcal{M}^p$. The minimax rate for the estimation of $\mathbf{E}(P)$ on $\mathcal{P}$ is

$$\mathcal{R}_n(\mathcal{P}) := \inf_{\hat{\mu}_n} \sup_{P \in \mathcal{P}} \mathbb{E}[\mathrm{OT}_p^p(\hat{\mu}_n, \mathbf{E}(P))], \quad (3.2)$$

where the infimum is taken over all possible estimators of $\mathbf{E}(P)$. An estimator attaining the rate $\mathcal{R}_n(\mathcal{P})$ (up to a constant) is called minimax, i.e. an estimator is minimax on the class $\mathcal{P}$ if it has the best possible risk uniformly on this class. We show that the empirical EPD $\overline{\mu}_n$ is a minimax estimator on $\mathcal{P}^q_{L,M}$ as long as $p \geq q + 1/2$. The case $p = \infty$ is discussed in Remark 1 (supplementary material).

Theorem 2. *Let $1 \leq p < \infty$ and $q \geq 0$, $L, M > 0$. One has, for some c depending on p and q,*

$$\mathcal{R}_n(\mathcal{P}^q_{L,M}) \geq cML^{p-q}n^{-1/2}. \quad (3.3)$$

As the EPD $\mathbf{E}(P)$ is known to have a smooth density in a wide variety of settings (Divol & Chazal, 2019), it could be expected (likewise it is the case in density estimation (Tsybakov, 2008)), that one could make use of this regularity to obtain substantially faster minimax rates on appropriate models. Surprisingly enough, using results from statistical optimal transport theory, we show that whatever regularity is assumed on the EPD, no estimators can perform better than the empirical EPD $\overline{\mu}_n$ for the OT_p loss (from a minimax perspective). Let $B^s_{p',q'}$ be the set of functions $\Omega \to \mathbb{R}$ in the Besov space of parameters $s \geq 0$ and $1 \leq p', q' \leq \infty$, see (Härdle et al., 2012) for an introduction to Besov spaces; note that this formalism encompasses all $\mathcal{C}^k$ classes. Consider the model $\mathcal{P}^{q,s}_{L,M,T}$ of probability distributions $P \in \mathcal{P}^q_{L,M}$ whose EPD $\mathbf{E}(P)$ belongs to $B^s_{p',q'}$ with associated norm smaller than T/M.

Theorem 3. *Let $1 \leq p < \infty$, $q, s \geq 0$, $L, M, T > 0$ and $1 \leq p', q' \leq \infty$. One has*

$$\mathcal{R}_n(\mathcal{P}^{q,s}_{L,M,T}) \geq cML^{p-q}n^{-1/2}, \quad (3.4)$$

where c depends on s, p', q', p, q and T.

The proof of Theorem 3 is based on a similar result appearing in (Weed & Berthet, 2019), where minimax rates of estimation with respect to the Wasserstein distance W_p are given for smooth densities on the cube.

Remark 1. *In the usual problem of estimating a measure thanks to a n-sample with respect to the Wasserstein distance, it has been noted several times (Trillos & Slepčev, 2015; Weed & Berthet, 2019; Divol, 2021) that this problem becomes significantly easier if the measure has a lower bounded density on its domain. In particular, it is known that the risk for the W_p^p loss of the empirical measure attains the faster rate $n^{-p/2}$ (instead of $n^{-1/2}$) under this hypothesis. If such a result is likely to hold for the OT_p^p loss under similar hypothesis, requiring that the EPD has a lower bounded density on some bounded domain U in Ω appears to be unreasonable. Indeed, this would imply that the density exhibits a sharp change of behavior at the boundary of U, whereas the density of the EPD is known to be typically smooth on Ω (Divol & Chazal, 2019). Whether there exists a more realistic assumption on the EPD for which the rate of convergence of the empirical EPD is $n^{-p/2}$ remains an open question.*

4. Quantization of the EPD

This section consists of two steps. In Section 4.1, we introduce and study the problem of quantizing persistence measures with respect to the metric OT_p, proving in particular the existence of optimal quantizers in general. Section 4.2 provides an online algorithm specifically designed to quantize EPD based on a sequence of observed diagrams $\mu_1, \ldots, \mu_n$ and provide theoretical guarantees of convergence.

4.1. Quantization for Persistence Measures

Let $\mu \in \mathcal{M}^p$ be a persistence measure and k be a fixed integer. The goal of the quantization problem is to build a

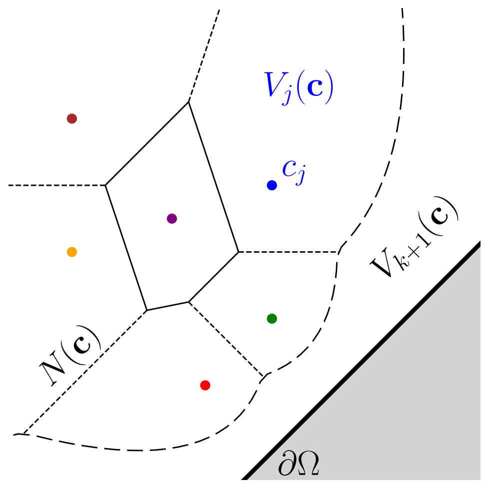

Figure 2. Example of partition $V_1(\mathbf{c}), \dots, V_{k+1}(\mathbf{c})$ for a given codebook $\mathbf{c}$.

measure $\nu = \sum_{j=1}^k m_j \delta_{c_j}$ supported on a set of k points $\mathbf{c} = (c_1, \dots, c_k)$ called a *codebook* (while the $(c_j)_j$s are called *centroids*) that approximates μ in an optimal way. Existing works (including previous works in the TDA literature) treat this problem over the space of probability measures equipped with the Wasserstein metric W_p over $\mathbb{R}^d$. Here, we use the metric OT_p instead, more suited to PDs, leading to benefits discussed in Remark 2 below. Our problem consists in minimizing the quantity $((m_1, c_1), \dots, (m_k, c_k)) \mapsto \mathrm{OT}_p\left(\sum_j m_j \delta_{c_j}, \mu\right)$ where $m_j \in \mathbb{R}_+$ and $c_j \in \Omega$. However, we show in Lemma 2 below that—as in the standard problem using the metric W_p—this problem can be reduced to an optimization problem on the codebook $\mathbf{c} \in \Omega^k$ only. To that aim, we introduce a notion of *Voronoï tesselation* relative to a codebook $\mathbf{c}$, with the subtlety that points closer to the diagonal $\partial\Omega$ define a specific cell, see Figure 2 for an illustration.

Definition 1. *Let $\mathbf{c} = (c_1 \dots c_k) \in \Omega^k$ and denote by convention $c_{k+1} := \partial\Omega$, so that in particular $\|x - c_{k+1}\| := \|x - \partial\Omega\|$. Define for $1 \leq j \leq k+1$,*

$$\begin{aligned} V_j(\mathbf{c}) :=& \{x \in \Omega,\ \forall j' < j, \|x - c_j\| \leq \|x - c_{j'}\| \\ & \text{and } \forall j' > j, \|x - c_j\| < \|x - c_{j'}\|\}, \\ N(\mathbf{c}) :=& \{x \in \Omega,\ \exists j < j' \text{ such that } x \in V_j(\mathbf{c}) \\ & \text{and } \|x - c_j\| = \|x - c_{j'}\|\}. \end{aligned} \tag{4.1}$$

Observe that $V_1(\mathbf{c}), \dots, V_{k+1}(\mathbf{c})$ form a partition of Ω.

Remark 2. *The difference between our approach and previous ones (in particular (Chazal et al., 2021)) lies in the presence of the "diagonal cell" $V_{k+1}(\mathbf{c})$. This cell introduces parabolic-shaped boundaries which slightly change the geometry of our problem. However, it has two major benefits. First, it enables a natural geometric identification of points close to the diagonal (which play a specific role in TDA) through the cell V_{k+1} and we do not "waste" centroids $(c_j)_{j=1}^k$ to encode them. Second, our approach does not require the introduction of a weight function (that artificially lowers the mass of points close to the diagonal), as typically done; removing the dependency on an important hyper-parameter.*

The following lemma states that given a persistence measure μ and a codebook $\mathbf{c} = (c_1, \dots, c_k)$, it is always optimal to set $m_j = \mu(V_j(\mathbf{c}))$.

Lemma 2. *Let $\mathbf{c} = (c_1, \dots, c_k)$. Let $\hat{\mu}(\mathbf{c}) := \sum_{j=1}^k \mu(V_j(\mathbf{c}))\delta_{c_j}$. Let $\nu = \sum_{j=1}^k m_j \delta_{c_j}$ for some $m_1, \dots, m_k \geq 0$. Then $\mathrm{OT}_p(\hat{\mu}(\mathbf{c}), \mu) \leq \mathrm{OT}_p(\nu, \mu)$.*

Therefore, quantizing μ boils down to the choice of the codebook $\mathbf{c}$. Formally, given a persistence measure μ to be quantized, a parameter $1 \leq p < \infty$ and an integer k, the quantization problem in the space of persistence measures consists in minimizing $R_{k,p} : \Omega^k \to \mathbb{R}$ defined for $\mathbf{c} \in \Omega^k$ by

$$\begin{aligned} R_{k,p}(\mathbf{c}) &:= \mathrm{OT}_p(\hat{\mu}(\mathbf{c}), \mu) \\ &= \left(\sum_{j=1}^{k+1} \int_{V_j(\mathbf{c})} \|x - c_j\|^p \mathrm{d}\mu(x) \right)^{\frac{1}{p}}, \end{aligned} \tag{4.2}$$

To alleviate notations, we write R_k instead of $R_{k,p}$ when the parameter p does not play a significant role. The value $R_k(\mathbf{c})$ is called the *distortion* achieved by $\mathbf{c}$. Let $R_k^* := \inf_{\mathbf{c} \in \Omega^k} R_k(\mathbf{c})$ and let $\mathbf{C}_k := \arg\min_{\mathbf{c} \in \Omega^k} R_k(\mathbf{c})$ be the set of optimal codebooks. Note that $R_k^* = 0$ if (and only if) $|\mathrm{spt}(\mu)| \leq k$. From now on, *we assume that μ has at least k points in its support.*

We can now state the main result of this subsection: the existence of an optimal codebook $\mathbf{c}^*$ for any persistence measure in $\mathcal{M}^p$. This result shares key ideas with (Graf & Luschgy, 2007, Thm 4.12), although we replace the assumption of finite p-th moment of the measure to be quantized by the assumption of finite total persistence $\mathrm{Pers}_p(\mu) < \infty$, more natural in TDA ($\mu$ may even have infinite total mass in our setting).

Proposition 4 (Existence of minimizers)**.** *The set of optimal codebooks $\mathbf{C}_k$ is a non-empty compact set. Furthermore, if $\mathbf{c}^* \in \mathbf{C}_k$, then, for all $1 \leq j \neq j' \leq k$, $\mu(V_j(\mathbf{c}^*)) > 0$ and $c_j^* \neq c_{j'}^*$.*

Corollary 1. *The following quantities are positive:*

$$\begin{aligned} D_{\min} &:= \inf_{\mathbf{c}^* \in \mathbf{C}_k, 1 \leq j \neq j' \leq k+1} \|c_j^* - c_{j'}^*\|, \\ m_{\min} &:= \inf_{\mathbf{c}^* \in \mathbf{C}_k, 1 \leq j \leq k} \mu(V_j(\mathbf{c}^*)). \end{aligned} \tag{4.3}$$

Computational aspects. One could consider to numerically solve the quantization problem (4.2) deriving optimization algorithms based on their counterpart in the optimal

Algorithm 1 Online quantization of EPDs

Input: A sequence $\mu_1, \dots, \mu_n$, integer k, parameter p.
Preprocess: Divide indices $\{1, \dots, n\}$ into batches $(B_1, \dots, B_T)$ of size $(n_1, \dots, n_T)$. Furthermore, divide $(B_t)_t$ into two halves $B_t^{(1)}$ and $B_t^{(2)}$.
Set $\overline{\mu}_t^{(\alpha)} := \frac{2}{n_t} \sum_{i \in B_t^{(\alpha)}} \mu_i$ for $1 \le t \le T, \alpha \in \{1, 2\}$.
Init: Sample $c_1^{(0)} \dots c_k^{(0)}$ from the diagrams.
for $t = 0, \dots, T-1$ **do**
 $\mathbf{c}^{(t+1)} = U_p(t, \mathbf{c}^{(t)}, \overline{\mu}_{t+1}^{(1)}, \overline{\mu}_{t+1}^{(2)})$ using (4.4)
end for
Output: The final codebook $\mathbf{c}^{(T)}$.

transport literature (Cuturi & Doucet, 2014), see (Lacombe, 2020, §7.2) for instance. However, using such techniques to quantize empirical EPDs would not be satisfactory for two reasons. First, the empirical EPD has in general a large number of points, hindering computational efficiency. Second, we want to leverage the fact that we observe a sequence of diagrams $\mu_1, \dots, \mu_n$, and not only their sum, to design an online algorithm that remains tractable with large sequences of large diagrams.

4.2. Quantization of an Empirical EPD

In Algorithm 1, we propose an online algorithm—adapted from (Chazal et al., 2021, Alg. 2) to the context of PDs and with arbitrary $p > 1$ instead of $p = 2$—that takes a sequence of observed PDs $\mu_1, \dots, \mu_n$ (a n-sample of law P) and outputs a codebook $(c_1, \dots, c_k)$ aiming at approximating $\mathbf{E}(P)$. The algorithm relies on an update function U_p for $p > 1$ defined as

$$U_p(t, \mathbf{c}, \mu, \mu') := \mathbf{c} - \frac{\left(\frac{\mu(V_j(\mathbf{c}))}{\mu'(V_j(\mathbf{c}))} \left(c_j - v_p(\mathbf{c}, \mu)_j\right)\right)_j}{t+1}, \quad (4.4)$$

where $v_p(\mathbf{c}, \mu)_j$ is the *p-center* of mass of μ over the cell $V_j(\mathbf{c})$:

$$v_p(\mathbf{c}, \mu)_j := \arg\min_y \left(\int_{V_j(\mathbf{c})} \|y - x\|^p \mathrm{d}\mu(x) \right)^{\frac{1}{p}}. \quad (4.5)$$

When $p = 2$, one simply has $v_2(\mathbf{c}, \mu)_j = \int_{V_j(\mathbf{c})} x \frac{\mathrm{d}\mu(x)}{\mu(V_j(\mathbf{c}))}$ and if in addition $\mu = \mu'$, the update (4.4) simplifies to

$$c_j \mapsto \frac{t}{t+1} c_j + \frac{1}{t+1} \int_{V_j(\mathbf{c})} x \frac{\mathrm{d}\mu(x)}{\mu(V_j(\mathbf{c}))},$$

so that roughly speaking, we are pushing c_j toward the usual center of mass of μ over the cell $V_j(\mathbf{c})$, similar to what is done when using the Lloyd algorithm to solve the k-means problem (Lloyd, 1982). More generally, (4.4) can be understood as pushing c_j toward the point that would decrease the distortion $R_{k,p}$ over the cell $V_j(\mathbf{c})$ the most, using a step-size (or learning rate) $\frac{1}{t+1}$. There is no closed-form for v_p for $p \neq 2$, though standard convex solvers may be used (Gonin, 1989). When $p = +\infty$, a central situation in TDA as it means working with the bottleneck distance OT_∞, computing v_∞ boils down to get the center of the *smallest enclosing circle* of $V_j(\mathbf{c}) \cap \mathrm{spt}(\mu)$. When μ is a discrete measure (e.g. an empirical EPD), this problem can be solved in linear time with respect to the number of points of μ that belong to $V_j(\mathbf{c})$ (Megiddo, 1983).

Note that in Algorithm 1, the split of batches $B_t = (B_t^{(1)}, B_t^{(2)})$ is only required for technical considerations (see the supplementary material and (Chazal et al., 2021)). In practice, this algorithm can be used without further assumptions and empirically, using $B_t = B_t^{(1)} = B_t^{(2)}$ yields substantially similar results. We provide a theoretical analysis of Algorithm 1 in the case $p = 2$, in particular through Theorem 5 which states that this algorithm is nearly optimal as a way to quantize $\mathbf{E}(P)$, provided the initialization is good enough. As in Section 3, we consider a probability distribution $P \in \mathcal{P}_{L,M}^p$. For $t > 0$ and $A \subset \Omega$, we let $A^t := \{x \in \Omega, \exists a \in A, \|x - a\| \le t\}$ be the t-neighborhood of A.

Definition 2 (Margin condition)**.** *Let $\mathbf{c}^*$ be an optimal quantizer of $\mathbf{E}(P)$. We say that P satisfies a* margin condition *of parameter $\lambda > 0$ and radius r_0 at $\mathbf{c}^*$ if, for all $t \in [0, r_0]$, one has $\mathbf{E}(P)(N(\mathbf{c}^*)^t) \le \lambda t$.*

Margin-like conditions on optimal codebook are standard in quantization literature (Tang & Monteleoni, 2016; Levrard, 2018). Informally, it indicates that the EPD concentrates around k poles, aside from the mass that is distributed close to the diagonal $\partial\Omega$; the smaller the λ, the more concentrated the measure. Note that this condition holds as long as the $\mathbf{E}(P)$ has a bounded density (although with possibly large λ), a property which is satisfied in a large number of situations, see (Divol & Chazal, 2019).

The following theorem states that given a n-sample of law P, Algorithm 1 outputs in $T = \frac{n}{\log(n)}$ steps a codebook $\mathbf{c}^{(T)}$ that approximates (in expectation) an optimal codebook $\mathbf{c}^*$ for $\mathbf{E}(P)$ at rate $\frac{\log(n)}{n}$, to be compared with the optimal rate of $\frac{1}{n}$ (Levrard, 2018, Prop. 7). It echoes (Chazal et al., 2021, Thm. 5) with the difference that, thanks to the diagonal cell V_{k+1}, we require a uniform bound on the total persistence of the measures rather than a uniform bound on their total mass, a more natural assumption in TDA.

Theorem 5. *Let $p = 2$. Let $P \in \mathcal{P}_{L,M}^2$ and let $\mathbf{c}^*$ be an optimal codebook for $\mathbf{E}(P)$. Assume that P satisfies a margin condition at $\mathbf{c}^*$ with parameters r_0 large enough and λ small enough (with respect to $D_{\min}, m_{\min}, L$ and M). Let $\mu_1, \dots, \mu_n$ be a n-sample of law P and $B_1, \dots, B_T$ be equally sized batches of length $C_1 \log(n)$. Finally, let $\mathbf{c}^{(T)}$ denote the output of Algorithm 1. There exists $R_0 > 0$ such*

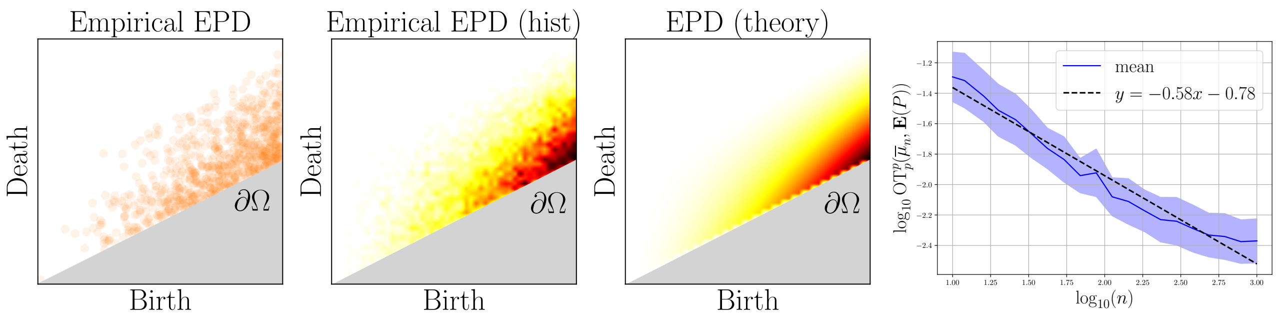

Figure 3. From left to right. (a) Empirical EPD $\overline{\mu}_n$ with $n = 10^3$. (b) Histogram of the empirical EPD on a 50×50 grid. (c) EPD $\mathbf{E}(P)$ of P, displayed on the same grid. (d) Distance $\mathrm{OT}_p^p(\overline{\mu}_n, \mathbf{E}(P))$ for $p = 2$ for different values of n in log-log scale (mean and standard deviation over 100 runs). A linear regression shows a convergence rate of order $n^{-0.58}$, close to the theoretical rate of $n^{-1/2}$ indicated by Theorem 1.

that if $\|\mathbf{c}^{(0)} - \mathbf{c}^*\| \leq R_0$, *then*

$$\mathbb{E}\|\mathbf{c}^{(T)} - \mathbf{c}^*\|^2 \leq C_2(\log n)/n,$$

where C_1, C_2 *and* R_0 *are constants depending on* $p, L, M, k, D_{\min}$ *and* $m_{\min}$.

5. Numerical Illustrations

We now provide some numerical illustrations that showcase our different theoretical results and their use in practice. Throughout, PDs are computed using the `Gudhi` library (Maria et al., 2014) and OT_p distances are computed building on tools available from the `POT` library (Flamary et al., 2021). See the supplementary material for further implementation details and complementary experiments.

Convergence rates for the empirical EPD. We first showcase the rate of convergence of Theorem 1. There are only few cases where explicit expressions for the EPD of a process are known. For instance, for Čech PDs based on a random sample of points, the corresponding EPD is known in closed-form only if the sample is supported on $\mathbb{R}$ (Divol & Polonik, 2019, Rem. 4.5). We therefore first consider a simple setting where an explicit expression can be derived. Let X be a set of N triangles $T_1, \ldots, T_N$, where N is uniform on $\{1, \ldots, 20\}$. We let $f : X \to \mathbb{R}$ be a random piecewise constant function, which is equal to $U_{i,j}$ on the jth edge of the triangle T_i, where the variables $(U_{i,j})$ are i.i.d. uniform variables on $[0, 1]$.

Furthermore, the function f is equal to $\max_{j=1,2,3} U_{i,j} + V_i$ on the inside of the triangle T_i, where the V_is are independent, independent from the $U_{i,j}$s, and follow a Beta distribution $\beta(1, 3)$. Let P be the distribution of the associated random PD. Let **rec** be the rectangle $[r_1, r_2] \times [s_1, s_2]$ for $r_1 \leq r_2 \leq s_1 \leq s_2$. Then,

$$\mathbf{E}(P)(\mathbf{rec}) = 30 \int_{r_1}^{r_2} t^2 \mathbb{P}(s_1 - t \leq V \leq s_2 - t) \mathrm{d}t, \quad (5.1)$$

where $V \sim \beta(1, 3)$. In practice, we compute $\mathbf{E}(P)$ on a discretization of $[0, 1] \times [0, 2]$ through a grid of size 50×50. Meanwhile, we sample empirical EPDs $\overline{\mu}_n$ for $10 \leq n \leq 10^3$. In order to estimate $\mathrm{OT}_p^p(\overline{\mu}_n, \mathbf{E}(P))$, we also turn these EPDs into histograms on the same grid, and then compute the OT_p distance between two histograms. See Figure 3 for an illustration which showcases in particular the expected rate $n^{-1/2}$.

We also exhibit the convergence of the empirical EPD in a more usual setting for the TDA practitioner. Namely, we build a random point cloud $\mathbf{X}$ with 10^3 points sampled on the surface of a torus with outer radius $r_1 = 5$ and inner radius $r_2 = 2$, and then consider the corresponding random Čech diagram for the 1-dimensional homology (loops, see Section 2). Given n realizations of $\mathbf{X}$, we compute the empirical EPD $\overline{\mu}_n$, where n ranges from 10 to $n_{max} = 1000$. As no closed-form for the corresponding EPD is known, we use as a proxy the empirical EPD based on a sample of size $2n_{\max}$, and then showcase in Figure 5 the convergence of $\mathrm{OT}_p^p(\overline{\mu}_n, \overline{\mu}_{2n_{\max}})$ at rate $n^{-1/2}$.

Quantization of the EPD. We now illustrate the behavior of Algorithm 1 using $p = 2$ and $p = \infty$ (referred to as "OT_2" and "OT_∞", respectively) and compare it to two natural alternatives. (Chazal et al., 2021, Alg. 2) is essentially the same algorithm without the "diagonal cell" $V_{k+1}(\mathbf{c})$; as such, centroids are dramatically influenced by points close to the diagonal which are likely to be abundant in standard applications of TDA. It is referred to as "W_2" in our illustrations, as it relies on quantization with respect to the Wasserstein distance with $p = 2$. The second alternative, referred to as "weighted codebook", is the one proposed in (Zieliński et al., 2020), which can be summarized in the following way: consider the empirical EPD $\overline{\mu}_n$ built on top of observations $\mu_1, \ldots, \mu_n$ (that is, concatenate the diagrams), and then subsample N points in the support of the empirical EPD, with the subtlety that the probability of choosing a point $x \in \mathrm{spt}(\overline{\mu}_n)$ depends on a

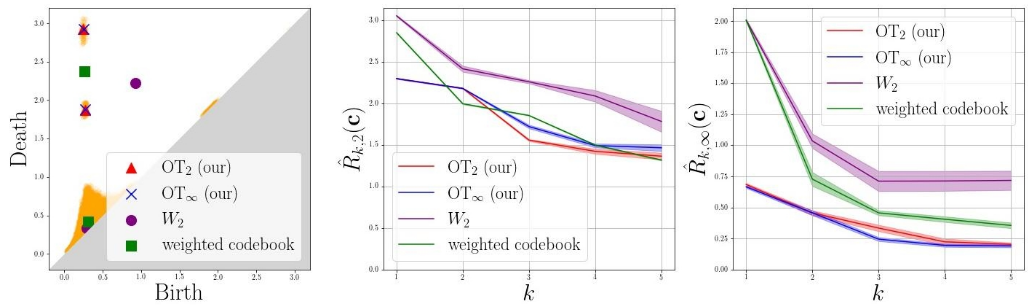

Figure 4. From left to right: (a) The quantization output for the different approaches considered with $k = 2$. As our approach accounts for the diagonal through the cell V_{k+1}, our codebooks retrieve the two clusters present in the EPD, while other approaches have one centroid used to account for the mass close to the diagonal. (b,c) The average distortion $R_{k,p}$ over 10 runs for the different methods, with $p = 2$ and $p = +\infty$.

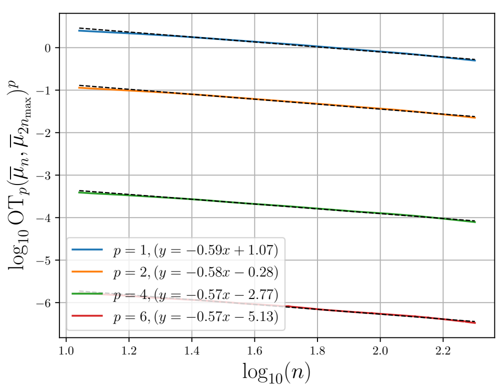

Figure 5. Convergence for $p = 1, 2, 4, 6$, each exhibiting a rate $\sim n^{-1/2}$.

weight function $w : \Omega \to \mathbb{R}_+$. Typical choices for w are of the form $w(x) = \min\left(\max\left(0, \frac{(\|x-\partial\Omega\|^q - \lambda)}{\theta - \lambda}\right), 1\right)$ for some parameters (λ, q, θ); the goal being to favor sampling points far from the diagonal. Zieliński et al. propose, in practice, to sample $N = 10^4$ points and to set $q = 1$, while λ and θ are the 0.05 and 0.95 quantiles of the distribution of $\{\|x - \partial\Omega\|^q,\ x \in \mathrm{spt}(\overline{\mu}_n)\}$, respectively. We use these parameters in our experiments. One then runs the Lloyd algorithm (k-means) on the set of N points that have been sampled to obtain a quantization of the empirical EPD.

We compare the different approaches in the following experiment. We randomly sample a point cloud $\mathbf{X}$ of size $\mathbf{m}$ on the surface of a torus with radii $(\mathbf{r_1}, \mathbf{r_2})$, where $\mathbf{m}, \mathbf{r_1}, \mathbf{r_2}$ are random variables that respectively follow a Poisson distribution of parameter $m \in \mathbb{N}$, a uniform distribution over $[r_1 - \varepsilon, r_1 + \varepsilon]$ and a uniform distribution over $[r_2 - \varepsilon, r_2 + \varepsilon]$. We use $m = 2,000, \varepsilon = 0.1, r_1 = 5$ and $r_2 = 2$ in our experiments. Given such a random point cloud $\mathbf{X}$, we build the Čech persistence diagram of its 1-dimensional features, denoted by μ, leading to a distribution P of PDs.

We then build a n-sample $\mu_1, \ldots, \mu_n$ with $n = 100$ and, for $k \in \{1, \ldots, 5\}$, compute the different codebooks returned by the aforementioned methods, using batches of size 10 for $\mathrm{OT}_2, \mathrm{OT}_\infty$ and W_2. All algorithms are initialized in the same way: we select the k points of highest persistence in the first diagram μ_1. To compare the quality of these codebooks, we evaluate their distortion (4.2) with $p = 2$ and $p = \infty$. As we do not have access to the true EPD $\mathbf{E}(P)$, we approximate this quantity through its empirical counterpart $\hat{R}_{k,p}(\mathbf{c}) := \left(\int_\Omega \min_{1 \le j \le c_{k+1}} \|x - c_j\|^p \mathrm{d}\overline{\mu}_n(x)\right)^{\frac{1}{p}}$, with $\hat{R}_{k,\infty}(\mathbf{c}) = \max_{x \in \mathrm{spt}(\overline{\mu}_n)} \min_j \|x - c_j\|$. Results are given in Figure 4. Interestingly, when $p = 2$ our approach is on a par with the weighted codebook approach, but becomes substantially better when evaluated with $p = \infty$, that is using the bottleneck distance which is the most natural metric to handle PDs.

We perform another experiment on the ORBIT5K dataset (Adams et al., 2017, §6.4.1), a benchmark dataset in TDA made of 5 classes with 1000 observations each (split into 70%/30% training/test) representing different dynamical systems, turned into PDs through Čech filtrations. For each class $i \in \{1, \ldots, 5\}$, we compute a 2-quantization $\nu^{(i)}$ using our OT_2 algorithm and a 3-quantization $\zeta^{(i)}$ using the standard W_2 approach as in (Chazal et al., 2021), i.e. without the diagonal cell V_{k+1} (but with an additional centroid). We then build two simple classifiers: the predicted class assigned to a test diagram μ is $\arg\min_i \{\mathrm{OT}_2(\mu, \nu^{(i)})\}$ (resp. $(\mu, \zeta^{(i)})$). Our OT_2 classifier achieves a decent test accuracy of 61%. Advanced (kernels, deep-learning) methods in TDA reach between 72% and 87% of accuracy (Carrière et al., 2020, Table 1); but we stress that our classifier is extremely simple (we summarize a whole training class by a measure with only $k = 2$ points!), showcasing that our quantizations summarize the training PDs in an informative way. More importantly, the W_2 classifier (with $k = 3$) only achieves 50% of test accuracy even though benefiting from

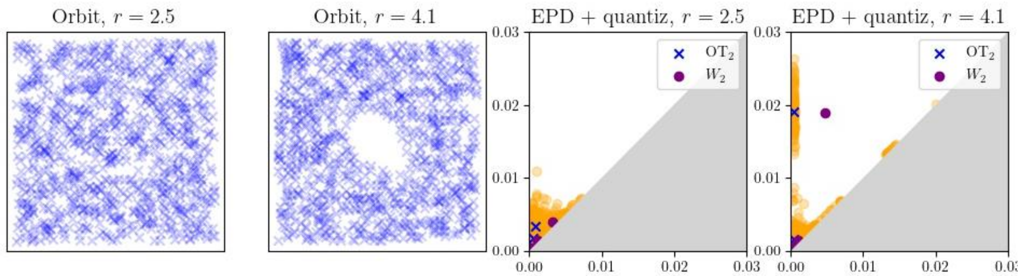

Figure 6. (Left) Two observations of the ORBIT5K dataset from two different classes (whose dynamics depend on a parameter r, see (Adams et al., 2017) for details). (Right) The empirical EPD (orange) observed for these two classes and the corresponding quantization obtained using our OT_2 algorithm with $k = 2$ and the W_2 algorithm (Chazal et al., 2021) with $k = 3$. As we account for the diagonal in a natural geometric way in our formulation, our quantization reflects the structure of the empirical EPD in a better way. This is especially striking in the case $r = 4.1$ (most right plot) where a centroid for the W_2 algorithm is deviated to a peculiar position due to the presence of few points close to the diagonal. Such points belong to the diagonal cell V_{k+1} in our setting.

an additional centroid, illustrating the importance of properly accounting for the diagonal as done in our approach.

6. Conclusion

This work is dedicated to the estimation of expected persistence diagrams, for which we prove that they are approximated, for the natural diagram metrics OT_p, by their empirical counterpart in an optimal way from a minimax perspective. We then introduce and study the quantization problem in the space of persistence diagrams, proving results of independent interest. Finally, we introduce an online algorithm to estimate a quantization of the EPD with theoretical guarantees. Interestingly, our algorithm can handle the case $p = \infty$, central in TDA, and has the advantage of not requiring hyper-parameters to account for the peculiar role played by the diagonal. We illustrate our results in numerical experiments and our code will be made publicly available. We believe that this work offers new perspectives to handle sample of PDs in practice and that it strengthens our understanding of statistical properties of PDs in random settings.

Aknowledgements. Authors thank Clément Levrard for thoughtful discussions.

References

Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F., and Ziegelmeier, L. Persistence images: a stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8):1–35, 2017.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Bourne, D. P., Schmitzer, B., and Wirth, B. Semi-discrete unbalanced optimal transport and quantization. *arXiv preprint arXiv:1808.01962*, 2018.

Bubenik, P. Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16(1):77–102, 2015.

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

Buchet, M., Hiraoka, Y., and Obayashi, I. Persistent homology and materials informatics. In *Nanoinformatics*, pp. 75–95. Springer, Singapore, 2018.

Cámara, P. G. Topological methods for genomics: present and future directions. *Current opinion in systems biology*, 1:95–101, 2017.

Carrière, M., Oudot, S. Y., and Ovsjanikov, M. Stable topological signatures for points on 3d shapes. In *Computer Graphics Forum*, volume 34, pp. 1–12. Wiley Online Library, 2015.

Carrière, M., Cuturi, M., and Oudot, S. Sliced Wasserstein kernel for persistence diagrams. In *34th International Conference on Machine Learning*, 2017.

Carrière, M., Chazal, F., Ike, Y., Lacombe, T., Royer, M., and Umeda, Y. Perslay: a neural network layer for persistence diagrams and new graph topological signatures. In *International Conference on Artificial Intelligence and Statistics*, pp. 2786–2796. PMLR, 2020.

Chazal, F., De Silva, V., Glisse, M., and Oudot, S. *The structure and stability of persistence modules*. Springer, 2016.

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

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.

Cohen-Steiner, D., Edelsbrunner, H., Harer, J., and Mileyko, Y. Lipschitz functions have L_p-stable persistence. *Foundations of computational mathematics*, 10(2):127–139, 2010.

Cuturi, M. and Doucet, A. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pp. 685–693, 2014.

Divol, V. A short proof on the rate of convergence of the empirical measure for the wasserstein distance. *arXiv preprint arXiv:2101.08126*, 2021.

Divol, V. and Chazal, F. The density of expected persistence diagrams and its kernel based estimation. *Journal of Computational Geometry*, 10(2):127–153, 2019.

Divol, V. and Lacombe, T. Understanding the topology and the geometry of the space of persistence diagrams via optimal partial transport. *Journal of Applied and Computational Topology*, pp. 1–53, 2020.

Divol, V. and Polonik, W. On the choice of weight functions for linear representations of persistence diagrams. *Journal of Applied and Computational Topology*, 3(3): 249–283, 2019.

Dudley, R. M. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1): 40–50, 1969.

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

Edelsbrunner, H., Letscher, D., and Zomorodian, A. Topological persistence and simplification. In *Proceedings 41st annual symposium on foundations of computer science*, pp. 454–463. IEEE, 2000.

Figalli, A. and Gigli, N. A new transportation distance between non-negative measures, with applications to gradients flows with dirichlet boundary conditions. *Journal de mathématiques pures et appliquées*, 94(2):107–130, 2010.

Fischer, A. Quantization and clustering with Bregman divergences. *Journal of Multivariate Analysis*, 101(9):2207–2221, 2010.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., et al. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Fournier, N. and Guillin, A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

Gonin, R. *Nonlinear Lp-norm estimation*, volume 100. CRC Press, 1989.

Graf, S. and Luschgy, H. *Foundations of quantization for probability distributions*. Springer, 2007.

Hardle, W., Park, B., and Tsybakov, A. Estimation of non-sharp support boundaries. *Journal of Multivariate Analysis*, 55(2):205–218, 1995.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. *Wavelets, approximation, and statistical applications*, volume 129. Springer Science & Business Media, 2012.

Kloeckner, B. R. Empirical measures: regularity is a counter-curse to dimensionality. *ESAIM: Probability and Statistics*, 24:408–434, 2020.

Lacombe, T. *Statistics for topological descriptors using optimal transport*. Theses, Institut Polytechnique de Paris, September 2020. URL https://hal.archives-ouvertes.fr/tel-02979251.

Lacombe, T., Cuturi, M., and Oudot, S. Large scale computation of means and clusters for persistence diagrams using optimal transport. In *Advances in Neural Information Processing Systems*, 2018.

Lei, J. et al. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767–798, 2020.

Levrard, C. Quantization/clustering: when and why does k-means work? *Journal de la Société Française de Statistique*, 159(1):1–26, 2018.

Levrard, C. et al. Nonasymptotic bounds for vector quantization in Hilbert spaces. *The Annals of Statistics*, 43(2): 592–619, 2015.

Li, C., Ovsjanikov, M., and Chazal, F. Persistence-based structural recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1995–2002, 2014.

Lloyd, S. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2):129–137, 1982.

Maria, C., Boissonnat, J.-D., Glisse, M., and Yvinec, M. The gudhi library: simplicial complexes and persistent homology. In *International Congress on Mathematical Software*, pp. 167–174. Springer, 2014.

Megiddo, N. Linear-time algorithms for linear programming in $\mathbb{R}^3$ and related problems. *SIAM journal on computing*, 12(4):759–776, 1983.

Reininghaus, J., Huber, S., Bauer, U., and Kwitt, R. A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4741–4748, 2015.

Saadatfar, M., Takeuchi, H., Robins, V., Francois, N., and Hiraoka, Y. Pore configuration landscape of granular crystallization. *Nature communications*, 8(1):1–11, 2017.

Santambrogio, F. Optimal transport for applied mathematicians. *Birkäuser, NY*, 2015.

Singh, S. and Póczos, B. Minimax distribution estimation in Wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.

Tang, C. and Monteleoni, C. On Lloyd's algorithm: new theoretical insights for clustering in practice. In *Artificial Intelligence and Statistics*, pp. 1280–1289. PMLR, 2016.

Trillos, N. G. and Slepčev, D. On the rate of convergence of empirical measures in ∞-transportation distance. *Canadian Journal of Mathematics*, 67(6):1358–1383, 2015.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519.

Turner, K., Mileyko, Y., Mukherjee, S., and Harer, J. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry*, 52(1):44–70, 2014.

Vidal, J., Budin, J., and Tierny, J. Progressive Wasserstein barycenters of persistence diagrams. *IEEE transactions on visualization and computer graphics*, 26(1):151–161, 2019.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Wagner, A. Nonembeddability of persistence diagrams with $p > 2$ Wasserstein metric. *Proceedings of the American Mathematical Society*, 149(6):2673–2677, 2021.

Weed, J. and Berthet, Q. Estimation of smooth densities in Wasserstein distance. In *Conference on Learning Theory*, pp. 3118–3119. PMLR, 2019.

Zhao, Q. and Wang, Y. Learning metrics for persistence-based summaries and applications for graph classification. *Advances in neural information processing systems (NeurIPS)*, 2019.

Zieliński, B., Lipiński, M., Juda, M., Zeppelzauer, M., and Dłotko, P. Persistence bag-of-words for topological data analysis. *arXiv preprint arXiv:1812.09245*, 2018.

Zieliński, B., Lipiński, M., Juda, M., Zeppelzauer, M., and Dłotko, P. Persistence codebooks for topological data analysis. *Artificial Intelligence Review*, pp. 1–41, 2020.

Zomorodian, A. and Carlsson, G. Computing persistent homology. *Discrete & Computational Geometry*, 33(2): 249–274, 2005.