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 successful 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., 2017; Buchet et al., 2018), cellular data (Cámara, 2017), 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 $\text{OT}_p$, 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 $\mu_1, \ldots, \mu_n$, 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 $(\mathcal{D}, \text{OT}_p)$ 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 $\text{OT}_p$ 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) $E(P)$ 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 $\Omega$ 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 $\mu_1, \ldots, \mu_n$ following some (unknown) law $P$. A natural way to estimate the EPD of $P$...
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 (Zielinski 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 $\Omega$ 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 (Zielinski et al., 2020)) but work in an online fashion with a sequence of observed diagrams. Second, we work with the standard diagram metric $\text{OT}_p$. 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 $\text{OT}_p$ 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

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

Related Work. Divol & Chazal (2019) show that under mild assumptions the EPD is a measure with density supported on the half-plane $\Omega$, 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 $\text{OT}_p$, 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 & Poczos, 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 $\text{OT}_p$ and the Wasserstein metric $W_p$ (see Section 2). Furthermore, we are not in the common situation where one observes i.i.d. realizations $X_1, \ldots, X_n$ in $\Omega$ 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 $\mu_1, \ldots, \mu_n$ on $\Omega$ following some law $P$ and considers the distance between the expected measure $\mathbb{E}(P)$ 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

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 $\mathbb{E}(P)$ at the parametric rate $n^{-1/2}$ with respect to the loss $\text{OT}_p$ 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.
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 result is 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 \( D_{\text{adm}}(f) := \sum_i \delta_{x_i} \), where \( \delta_{x_i} \) denotes the Dirac mass located at \( x_i \in \Omega \).

Of particular interest is the case where \( X = \mathbb{R}^d \), and \( f : w \in \mathbb{R}^d \mapsto \text{dist}(w, A) \) 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 \( D_{\text{adm}}(A) \).

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

\[
\text{OT}_p(\mu, \nu) := \inf_{\pi \in \text{Adm}(\mu, \nu)} \left( \int_{\Omega \times \Omega} \|x-y\|^p \, d\pi \right)^{\frac{1}{p}}, \tag{2.1}
\]

where \( \text{Adm}(\mu, \nu) \) is the set of measures supported on \( \Omega \times \Omega \) whose first (resp. second) marginal coincides with \( \mu \) (resp. \( \nu \)) on \( \Omega \) (note in particular that \( \pi \) is not constrained on \( \partial \Omega \times \partial \Omega \)). The definition is extended to \( p = \infty \) by replacing \( \left( \int_{\Omega \times \Omega} \|x-y\|^p \, d\pi \right)^{-\frac{1}{p}} \) by \( \sup \{ \|x-y\|, (x, y) \in \text{spt}(\pi) \} \), and the distance \( \text{OT}_\infty \) 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 \( (\mathcal{M}_p, \text{OT}_p) \) of persistence measures is defined as the space of (non-negative) Radon measures \( \mu \) supported on \( \Omega \) that have finite total persistence, i.e. \( \text{Pers}_p(\mu) := \int \|x - \partial \Omega\|^p \, d\mu(x) < \infty \) (this condition ensures that \( \text{OT}_p(\mu) \) is always finite). Note that the distance \( \text{OT}_p \) is not only defined for PDs (elements of \( \mathcal{D} \)), but for measures on \( \Omega \) 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 \( \text{OT}_p \) are similar to the Wasserstein distances used in optimal transport (Santambrogio, 2015, Ch. 5): for \( \sigma, \tau \) two measures having the same total mass on a metric space \((S, \rho)\), the distance \( W_{p,\rho}(\sigma, \tau) \) is defined as the infimum of \( \left( \int_{S^2} \rho(x, y)^p \, d\pi(x, y) \right)^{1/p} \) over all transport plans \( \pi \) between \( \sigma \) and \( \tau \), i.e. measures on \( S \times S \) which have for first (resp. second) marginal \( \sigma \) (resp. \( \tau \)). When \( p \) 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 \( \text{OT}_p \) distance: the constraints in (2.1) only involves the marginals on \( \Omega \), allowing us to transport mass to and from the boundary of the space \( \partial \Omega \). It makes, in particular, the distance \( \text{OT}_p \) between measures of different total masses well-defined. The metrics \( \text{OT}_p \) 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 \((\mathcal{M}_p, \text{OT}_p)\). Let \( E(P) \) be the measure defined by, for \( A \subset \Omega \) compact,

\[
E(P)(A) := E_P[\mu(A)], \tag{2.2}
\]

where \( \mu \sim P \), and \( \mu(A) \) is the (random) number of points of \( \mu \) 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 \( \Omega \). 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 \( E(P) \) be its EPD. Given a \( n \)-sample \( \mu_1, \ldots, \mu_n \) of law \( P \), the empirical EPD is defined as \( \overline{\text{OT}}_p := \frac{1}{n} \sum \mu_i \). In this section, we control the distance \( \text{OT}_p^2(\overline{\text{OT}}_p, E(P)) \) under moment assumptions on the underlying law \( P \). Note that, according to (Divol & Lacombe, 2020, Thm. 3.7) and the law of large numbers, \( \overline{\text{OT}}_p \rightarrow_{OT} E(P) \) almost surely under the minimal assumption that \( E_P[\text{Pers}_p(\mu)] < \infty \) (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 \( \ell_1 \)-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 \( \text{Pers}_q(\mu) \leq M \). Let \( P^q_{L,M} \) 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
every $q > d$, $\text{Dgm}(f)$ belongs to $M_{L,M}^q$ 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 $\text{Pers}_n$ 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 $P \in \mathcal{P}_{L,M}^q$ and let $\mu_1, \ldots, \mu_n$ be a $n$-sample from law $P$. If $\pi_n$ is the associated empirical EPD, then

$$E[\text{OT}_p^q(\pi_n, E(P))] \leq c ML^{p-q} \left(\frac{1}{n^{1/2}} + \frac{a_p(n)}{n^{p-q}}\right), \quad (3.1)$$

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 $\mu_i = \text{Dgm}(f_i)$ 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 $E(P)$ exist. A possible way to answer this question is given by the minimax framework. Let $\mathcal{P}$ be a set of probability distributions on $\mathbb{R}^d$. The minimax rate for the estimation of $E(P)$ on $\mathcal{P}$ is

$$R_n(\mathcal{P}) := \inf_{\hat{\mu}_n} \sup_{P \in \mathcal{P}} E[\text{OT}_p^q(\hat{\mu}_n, E(P))], \quad (3.2)$$

where the infimum is taken over all possible estimators of $E(P)$. An estimator attaining the rate $R_n(\mathcal{P})$ (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 $\pi_n$ is a minimax estimator on $\mathcal{P}_{L,M}^q$ 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 $q$,

$$R_n(\mathcal{P}_{L,M}^q) \geq c ML^{p-q}n^{-1/2}. \quad (3.3)$$

As the EPD $E(P)$ 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 $\pi_n$ for the OT$_p$ loss (from a minimax perspective). Let $P_{q,p}^{q',q'}$ 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 $C^k$ classes. Consider the model $\mathcal{P}_{L,M,T}^{q,s}$ of probability distributions $P \in \mathcal{P}_{L,M}^q$ whose EPD $E(P)$ belongs to $B_{q,q}^{s,q'}$ 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

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

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 $n^{-p/p}$ (instead of $n^{-1/2}$) under this hypothesis. If such a result is likely to hold for the OT$_p$ loss under similar hypothesis, requiring that the EPD has a lower bounded density on some bounded domain $U$ in $\Omega$ 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 $\Omega$ (Divol & Chazal, 2019). Whether there exists a more realistic assumption on the EPD for which the rate of convergence of the empirical EPD is $n^{-p/p}$ 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
Remark 2. The difference between our approach and previous ones (in particular (Chazal et al., 2021)) lies in the presence of the “diagonal cell” \( V_{k+1}(c) \). 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 \( \nu \) and a codebook \( c = (c_1, \ldots, c_k) \), it is always optimal to set \( m_j = \mu(V_j(c)) \).

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

Therefore, quantizing \( \nu \) boils down to the choice of the codebook \( c \). Formally, given a persistence measure \( \mu \) 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 \( R_{k,p} : \Omega^k \rightarrow \mathbb{R} \) defined for \( c \in \Omega^k \) by

\[
R_{k,p}(c) := \text{OT}_p(\hat{\mu}(c), \mu)
\]

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

We can now state the main result of this subsection: the existence of an optimal codebook \( c^* \) for any persistence measure in \( M^p \). 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 \( \text{Pers}_p(\mu) < \infty \), more natural in TDA (\( \mu \) may even have infinite total mass in our setting).

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

Corollary 1. The following quantities are positive:

\[
D_{\text{min}} := \inf_{c^* \in C_k, 1 \leq j \neq j' \leq k+1} \|c_j^* - c_{j'}^*\|,
\]

\[
m_{\text{min}} := \inf_{c^* \in C_k, 1 \leq j \leq k} \mu(V_j(c^*)).
\]

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

Figure 2. Example of partition \( V_1(c), \ldots, V_{k+1}(c) \) for a given codebook \( c \).
Algorithm 1 Online quantization of EPDs

**Input:** A sequence $\mu_1, \ldots, \mu_n$, integer $k$, parameter $p$.

**Preprocess:** Divide indices $\{1, \ldots, n\}$ into batches $\{(B_t^1, B_t^2)\}$ of size $\{n_1, \ldots, n_T\}$. Furthermore, divide $(B_t)$ into two halves $B_t^{(1)}$ and $B_t^{(2)}$.

Set $\mu_t^{(\alpha)} := \frac{2}{n_t} \sum_{t \in B_t^{(\alpha)}} \mu_t$ for $1 \leq t \leq T$. Set $\alpha \in \{1, 2\}$.

**Init:** Sample $c_1^{(1)} \ldots c_k^{(1)}$ from the diagonals.

for $t = 0, \ldots, T - 1$ do

$c_{(t+1)}^{(1)} = U_p(t, c^{(t)}, \mu_t^{(1)}, \mu_t^{(2)})$ using (4.4)

end for

**Output:** The final codebook $c^{(T)}$.

Theorem 5. Let $p \geq 2$, then for large enough $\lambda$ and $\lambda$ small enough (with respect to $\Lambda_{\text{size}}$ and $\Lambda_{\text{scale}}$, the margin condition at $\lambda$, 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 $\lambda$), 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 $c^{(T)}$ that approximates (in expectation) an optimal codebook $c^*$ 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 $P \in \mathcal{P}_{\mathcal{L}, \mathcal{M}}$ and let $c^*$ be an optimal codebook for $\mathbf{E}(P)$. Assume that $P$ satisfies a margin condition at $c^*$ with parameters $r_0$ large enough and $\lambda$ small enough (with respect to $D_{\text{size}}, m_{\text{min}}, L$ and $M$). Let $\mu_1, \ldots, \mu_n$ be a $n$-sample of law $P$ and $B_1, \ldots, B_T$ be equally sized batches of length $C_1 \log(n)$. Finally, let $c^{(T)}$ denote the output of Algorithm 1. There exists $R_0 > 0$ such...
that if \( \|c^{(0)} - c^*\| \leq R_0 \), then
\[
E\|c^{(P)} - c^*\|^2 \leq C_2(\log n)/n,
\]
where \( C_1, C_2 \) and \( R_0 \) are constants depending on \( p, I, M, k, D_{\min} \) 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 treated using 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 a 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 \( j \)th 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 \( \max_{i,j=1,2,3} U_{i,j} + V_i \) on the inside of the triangle \( T_i \), where the \( V_i \)s 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 \( \text{rec} \) 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,
\[
E(P)(\text{rec}) = 30 \int_{r_1}^{r_2} t^2 \mathbb{P}(s_1 - t \leq V \leq s_2 - t)dt, \quad (5.1)
\]
where \( V \sim \beta(1, 3) \). In practice, we compute \( E(P) \) on a discretization of \([0, 1] \times [0, 2]\) through a grid of size \( 50 \times 50 \). Meanwhile, we sample empirical PDs \( \pi_n \) for \( 10 \leq n \leq 10^3 \). In order to estimate \( OT^p_{{\mathcal{P}}}(\pi_n, E(P)) \), we also turn these PDs 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 \( X \) 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 \( X \), we compute the empirical PD \( \pi_n \), where \( n \) ranges from 10 to \( 10^4 \). 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 \( OT^p_{{\mathcal{P}}}(\pi_n, \pi_{2n_{\max}}) \) at rate \( n^{-1/2} \).

#### Quantization of the EPD

We now illustrate the behavior of Algorithm 1 using \( p = 2 \) and \( \rho = \infty \) (referred to as “\( OT_2 \)” and “\( OT_{\infty} \)”, respectively) and compare it to two natural alternatives. (Chazal et al., 2021, Alg. 2) is essentially the same algorithm without the “diagonal cell” \( V_{k+1}(c) \); 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 PD \( \pi_n \) 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 PD, with the subtlety that the probability of choosing a point \( x \in \text{spt}(\pi_n) \) depends on a
We compare the different approaches in the following experiments. Given such a random point cloud \((\text{the surface of a torus with radii } r_1, r_2)\), we randomly sample a point cloud \(n\) points and to set \(q = 1\), while \(\lambda\) and \(\theta\) are the 0.05 and 0.95 quantiles of the distribution of \(\|x - \partial \Omega\| q, x \in \text{spt}(\pi_n)\), respectively. We use these parameters in our experiments. One then runs the Lloyd algorithm \((k\text{-means})\) 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 \(X\) of size \(m\) on the surface of a torus with radii \((r_1, r_2)\), where \(r_1, r_2\) 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 \(X\), we build the \(\check{\text{C}}\)ech persistence diagram of its 1-dimensional features, denoted by \(\mu\), 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 \(\text{OT}_2, \text{OT}_\infty\) and \(W_2\). All algorithms are initialized in the same way: we select the \(k\) points of highest persistence in the first diagram \(\mu_1\). 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 \(E(P)\), we approximate this quantity through its empirical counterpart \(\hat{R}_{k,p}(c) := (\int \min_{1 \leq j \leq c} \|x - c_j\|^p d\pi_n(x))^{\frac{1}{p}}\), with \(\hat{R}_{k,\infty}(c) = \max_{x \in \text{spt}(\pi_n)} \min_j \|x - c_j\|\). 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 \(\check{\text{C}}\)ech filtrations. For each class \(i \in \{1, \ldots, 5\}\), we compute a 2-quantization \(\nu^{(i)}\) using our \(\text{OT}_2\) 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 \(\mu\) is \(\arg \min_{i} \{\text{OT}_2(\mu, \nu^{(i)})\}\) (resp. \(\mu, \zeta^{(i)}\)). Our \(\text{OT}_2\) 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.

Acknowledgements. Authors thank Clément Levrard for thoughtful discussions.

References


Estimation and Quantization of EPDs


