Optimal quantization of the mean measure and application to clustering of measures

Frédéric Chazal, Clément Levrard, Martin Royer

Optimal quantization of the mean measure and application to clustering of measures

Frédéric Chazal [1], **Clément Levrard**[2] **and Martin Royer**[1]

[1] *DataShape, Inria Saclay, e-mail:* frederic.chazal@inria.fr; martin.royer@inria.fr

[2] *LPSM, Université de Paris, e-mail:* clement.levrard@lpsm.paris

Abstract: This paper addresses the case where data come as point sets, or more generally as discrete measures. Our motivation is twofold: first we intend to approximate with a compactly supported measure the mean of the measure generating process, that coincides with the intensity measure in the point process framework, or with the expected persistence diagram in the framework of persistence-based topological data analysis. To this aim we provide two algorithms that we prove almost minimax optimal.

Second we build from the estimator of the mean measure a vectorization map, that sends every measure into a finite-dimensional Euclidean space, and investigate its properties through a clustering-oriented lens. In a nutshell, we show that in a mixture of measure generating process, our technique yields a representation in $\mathbb{R}^k$, for $k \in \mathbb{N}^*$ that guarantees a good clustering of the data points with high probability. Interestingly, our results apply in the framework of persistence-based shape classification via the ATOL procedure described in [28].

1. Introduction

This paper handles the case where we observe n i.i.d measures $X_1, \ldots, X_n$, rather than n i.i.d sample points, the latter case being the standard input of many machine learning algorithms. Such kind of observations naturally arise in many situations, for instance when data are spatial point patterns: in species distribution modeling [27], repartition of clusters of diseases [13], modelisation of crime repartition [29] to name a few. The framework of i.i.d sample measures also encompasses analysis of multi-channel time series, for instance in embankment dam anomaly detection from piezometers [17], as well as topological data analysis via persistence diagrams [14, 7]. The objective of the paper is twofold: first we want to build from data a compact representation of the mean of the measure, in the arithmetic sense. Second, based on the first construction, we intend to provide a provably efficient clustering technique for measures.

Applications for the first objective might be found whenever the sample measures are organized around a central measure of interest, for instance in image analysis [12] or point processes modeling [27, 13, 29]. In [12], the central measure is defined as the Wasserstein barycenter of the distribution of measures. Namely, if we assume that $X_1, \ldots, X_n$ are i.i.d measures on $\mathbb{R}^d$ drawn from X, where X is a probability distribution on the space of measures, then the central

measure is defined as $\mu_W = \arg\min_\nu \mathbb{E}\left(W_2(X,\nu)\right)^2$, where ν ranges in the space of measures and W_2 denotes the Wasserstein distance. Note that this definition only makes sense in the case where $X(\mathbb{R}^d)$ is constant a.s., that is when we draw measures with the same total mass. Moreover, computing the Wasserstein barycenter of $X_1, \dots, X_n$ in practice is too costly for large n's, even with approximating algorithms [12, 25]. To overcome these difficulties, we choose to define the central measure as the arithmetic mean of X, denoted by $\mathbb{E}(X)$, that assigns the weight $\mathbb{E}\left[X(A)\right]$ to a borelian set A. In the point process theory, the mean measure is often referred to as the intensity function of the process.

An easily computable estimator of this mean measure is the sample mean measure $\bar{X}_n = \left(\sum_{i=1}^n X_i\right)/n$. We intend to build a k-points approximation of $\mathbb{E}(X)$, that is a distribution $P_{\mathbf{c}}$ supported by $\mathbf{c} = (c_1, \dots, c_k)$ that approximates well $\mathbb{E}(X)$, based on $X_1, \dots, X_n$. To this aim, we introduce two algorithms (batch and mini-batch) that extend classical quantization techniques intended to solve the k-means problem [22]. In fact, these algorithms are build to solve the k-means problem for $\bar{X}_n$. We prove in Section 2.2 that these algorithms provide minimax optimal estimators of a best possible k-points approximation of $\mathbb{E}(X)$, provided that $\mathbb{E}(X)$ satisfies some structural assumption. Interestingly, our results also proves optimality of the classical quantization techniques [22, 21] in the point sample case.

The second objective, clustering of measures, has a wide range of possible applications: in the case where data come as a collection of finite point sets for instance, including ecology [27], genetics [28, 2], graphs clustering [6, 16] and shapes clustering [7]. Our technique is based on a vectorization of the measures, that is a map v that sends every measure X_1 into $\mathbb{R}^k$. We build this vectorization using the optimal k-points $\mathbf{c} = (c_1, \dots, c_k)$ obtained in the first part (Section 2.2), transforming each X_i into a vector $v_i \in \mathbb{R}^k$ that roughly encodes how much weight X_i spreads around every c_j. Note that a vectorization based on a fixed grid of $\mathbb{R}^d$ is possible, however the dimension of such a vectorization would be quite large. In the particular framework of topological data analysis and persistence diagrams clustering, vectorization via evaluation onto a fixed grid is the technique exposed in [3], whereas our method has clear connections with the procedures described in [32, 28].

For this vectorization scheme, we provide general conditions on the structure of the sample measures that allow an almost exact clustering based on the vectorization space. It is worth mentioning that our theoretical results include vectorization via evaluations of kernel functions around each point c_j, for a general class of kernel functions that encompasses the one used in [28]. Further, we also prove in Section 4 that theses structural conditions are fulfilled in a framework of shape classification via persistence diagrams. As a consequence, we theoretically asses the performance of the procedure exposed in [28]. Up to our knowledge, this provides the only theoretical guarantee on such a persistence-based clustering algorithm.

The paper is organized as follows: in Section 2, we introduce notation along with the exposition of the problem of mean measure quantization. Then, two theoretically grounded algorithms are described to solve this problem from the

sample $X_1, \dots, X_n$. Section 3 exposes our general vectorization technique, and conditions that guarantee a correct clustering based on it. Section 4 investigates the special case where the measures are persistence diagrams built from samplings of different shapes, showing that all the previously exposed theoretical results apply in this framework. Sections 5, 6 and 7 gather the main proofs of the results. At last, Section 8 gives the proof of intermediate and technical results.

2. Quantization of the mean measure

2.1. Definition and notation

Throughout the paper we will consider finite measures on the d-dimensional ball $\mathcal{B}(0,R)$ of the Euclidean space $\mathbb{R}^d$, and denote by $\mathcal{M}(R,M)$ the set of such measures of total mass smaller than M. For an element $\mu \in \mathcal{M}(R,M)$ we denote by $M(\mu)$ its total mass. Further, if $\mu \in \mathcal{M}(R,M)$ and f is a borelian function from $\mathbb{R}^d$ to $\mathbb{R}$, we denote by $\mu(du) \bullet f(u)$ integration of f with respect to μ, whenever $\mu(du) \bullet |f|(u)$ is finite. We let X denote a random variable taking values in $\mathcal{M}(R,M)$, and $X_1, \dots, X_n$ denote an i.i.d. sample with the same distribution as X. Definition 1 below introduces the mean measure.

Definition 1. *Let $\mathcal{B}(\mathbb{R}^d)$ denote the borelian sets of $\mathbb{R}^d$. The mean measure $\mathbb{E}(X)$ is defined as the measure such that*

$$\forall A \in \mathcal{B}(\mathbb{R}^d) \quad \mathbb{E}(X)(A) = \mathbb{E}\left(X(A)\right).$$

As well, the empirical mean measure $\bar{X}_n$ may be defined via

$$\forall A \in \mathcal{B}(\mathbb{R}^d) \quad \bar{X}_n(A) = \frac{1}{n}\sum_{i=1}^{n} X_i(A).$$

In the case where the measures of interest are persistence diagrams, the mean measure defined above is the expected persistence diagram, defined in [10]. If the sample measures are point processes, $\mathbb{E}(X)$ is the intensity function of the process. It is straightforward that, if $\mathbb{P}\left(X \in \mathcal{M}(R,M)\right) = 1$, then both $\mathbb{E}(X)$ and $\bar{X}_n$ are (almost surely) elements of $\mathcal{M}(R,M)$. The goal of this paper is to build a k-points approximation of $\mathbb{E}(X)$ based on $X_1, \dots, X_n$.

If μ_1, $\mu_2 \in \mathcal{M}(R,M)$ satisfy $M(\mu_1) = M(\mu_2)$, and $p \in [\![1, +\infty]\!]$, we may define $W_p(\mu_1, \mu_2)$ as the p-Wasserstein distance between μ_1 and μ_2. Let $\mathcal{M}_k(R,M)$ denote the subset of $\mathcal{M}(R,M)$ that consists of distributions supported by k points. Adopting the vocabulary of quantization, each support point of a finite k-points distribution is called a codepoint, and the vector made of k codepoints $c_1, \dots, c_k$ is called a codebook. For any codebook $\mathbf{c} = (c_1, \dots, c_k) \in \mathcal{B}(0,R)^k$, we let

$$W_j(\mathbf{c}) = \{x \in \mathbb{R}^d \mid \forall i < j \quad \|x - c_j\| < \|x - c_i\| \quad \text{and} \quad \forall i > j \quad \|x - c_j\| \leq \|x - c_i\|\},$$
$$N(\mathbf{c}) = \{x \mid \exists i < j \quad x \in W_i(\mathbf{c}) \quad \text{and} \quad \|x - c_j\| = \|x - c_j\|\},$$

so that $(W_1(\mathbf{c}), \dots, W_k(\mathbf{c}))$ forms a partition of $\mathbb{R}^d$ and $N(\mathbf{c})$ represents the skeleton of the Voronoi diagram associated with $\mathbf{c}$. Given a codebook $\mathbf{c}$, a standard way to approximate $\mathbb{E}(X)$ with a probability distribution supported by $\mathbf{c}$ is to consider $P_{\mathbf{c}} = \sum_{j=1}^{k} \mathbb{E}(X)(W_j(\mathbf{c}))\delta_{c_j}$. It is then easy to see that, for any other distribution $P'_k = \sum_{j=1}^{k} \mu_j \delta_{c_j}$ such that $\sum_{j=1}^{k} \mu_j = M(\mathbb{E}(X))$, supported by $\mathbf{c}$,

$$W_2^2(\mathbb{E}(X), P'_k) \geq W_2^2(\mathbb{E}(X), P_{\mathbf{c}}) = \mathbb{E}(X)(du) \bullet \min_{j=1,\dots,k} \|u - c_j\|^2 = R(\mathbf{c}).$$

Thus, finding the best k-points approximation of $\mathbb{E}(X)$ in terms of W_2 boils down to minimize $R(\mathbf{c})$. Note that $R(\mathbf{c})$ is often referred to as the distortion of $\mathbf{c}$, in the quantization framework. According to [15, Corollary 3.1], since $\mathbb{E}(X) \in \mathcal{M}(R, M)$, there exists minimizers $\mathbf{c}^*$ of $R(\mathbf{c})$, and we let $\mathcal{C}_{opt}$ denote the set of such minimizers. In what follows, R^* will denote the optimal distortion achievable with k points, that is $R^* = R(\mathbf{c})$, where $\mathbf{c} \in \mathcal{C}_{opt}$. Basic properties of $\mathcal{C}_{opt}$ are recalled below.

Proposition 2. *[20, Proposition 1] Recall that $\mathbb{E}(X) \in \mathcal{M}(R, M)$, then*

1. $B = \inf_{\mathbf{c} \in \mathcal{C}_{opt}, j \neq i} \|c_i^* - c_j^*\| > 0$,
2. $p_{min} = \inf_{\mathbf{c} \in \mathcal{C}_{opt}, j=1,\dots,k} \mathbb{E}(X)\left(W_j(\mathbf{c}^*)\right) > 0$.

In what follows, we will further assume that $\mathbb{E}(X)$ satisfies a so-called *margin condition*, defined in [18, Definition 2.1] and recalled below.

Definition 3. *$\mathbb{E}(X) \in \mathcal{M}(R, M)$ satisfies a margin condition with radius $r_0 > 0$ if and only if, for all $0 \leq t \leq r_0$,*

$$\sup_{\mathbf{c}^* \in \mathcal{C}_{opt}} \mathbb{E}(X)\left(\mathcal{B}(N(\mathbf{c}^*), t)\right) \leq \frac{B p_{min}}{128 R^2} t.$$

In a nutshell, a margin condition ensures that the mean distribution $\mathbb{E}(X)$ is well-concentrated around k poles. Following [20], a margin condition will ensure that usual k-means type algorithms are almost optimal in terms of distortion. These algorithms are recalled below and adapted to the mean-measure quantization framework.

2.2. Batch and mini-batch algorithms

Let $X_1, \dots, X_n$ be i.i.d random measures in $\mathcal{M}_{N_{max}}(R, M)$ (where we recall that $\mathcal{M}_{N_{max}}(R, M)$ is the set of distributions in $\mathcal{M}(R, M)$ supported by at most N_{max} points). This section exposes two algorithms that are intended to approximate a best k-points empirical codebook, that is a codebook $\hat{\mathbf{c}}_n$ which minimizes $W_2(\bar{X}_n, \hat{P}_{\mathbf{c}})$, for $\mathbf{c} \in \mathcal{B}(0, R)^k$, $\hat{P}_{\mathbf{c}}$ being defined by $\sum_{j=1}^{k} \bar{X}_n(W_j(\mathbf{c}))\delta_{c_j}$. These algorithms are extensions of two well-known clustering algorithms, namely the Lloyd algorithm ([21]) and Mac Queen algorithm ([22]). We first introduce the counterpart of Lloyd algorithm.

Algorithm 1: Batch algorithm (Lloyd)

Input : $X_1, \ldots, X_n$ and k ;
 # Initialization
Sample $c_1^{(0)}, c_2^{(0)}, \ldots c_k^{(0)}$ from $\bar{X}_n$. ;
while $\mathbf{c}^{(t+1)} \neq \mathbf{c}^{(t)}$ do :
  # Centroid update.
  for j **in** $1..k$:
    $c_j^{(t+1)} = \frac{1}{\bar{X}_n(W_j(\mathbf{c}^{(t)}))} \bar{X}_n(du) \bullet \left[u \mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u) \right]$;
Output : $\mathbf{c}^{(T)}$ (codebook of the last iteration) .

Note that Algorithm 1 is a batch algorithm, in the sense that every iteration need to process the whole data set $X_1, \ldots, X_n$. Fortunately, Theorem 4 below ensures that a limited number of iterations are required for Algorithm 1 to provide an almost optimal solution. In the sample point case, that is when we observe n i.i.d points $X_i^{(1)}$, Algorithm 1 is the usual Lloyd's algorithm. In this case, the mean measure $\mathbb{E}(X)$ is the distribution of $X_1^{(1)}$, that is the usual sampling distribution of the n i.i.d points. As well, the counterpart of Mac-Queen algorithm ([22]) for standard k-means clustering is the following mini-batch algorithm.

Algorithm 2: Mini-batch algorithm (Mac-Queen)

Input : $X_1, \ldots, X_n$, divided into mini-batches $(B_1, \ldots, B_T)$ of sizes $(n_1, \ldots, n_T)$, and k ;
 # Initialization
Sample $c_1^{(0)}, c_2^{(0)}, \ldots c_k^{(0)}$ from $\bar{X}_n$. ;
for $j = 0, \ldots, T-1$ do :
  # Centroid update.
  for j **in** $1..k$:
    $c_j^{(t+1)} = c_j^{(t)} - \frac{1}{(t+1)\bar{X}_{B_{t+1}}\left(W_j(\mathbf{c}^{(t)})\right)} \bar{X}_{B_{t+1}}(du) \bullet \left[(c_j^{(t)} - u) \mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u) \right]$;
Output : $\mathbf{c}^{(T)}$ (codebook of the last iteration) .

Whenever $n_i = 1$ for $i = 1, \ldots, n$, Algorithm 2 is a slight modification of the original Mac-Queen algorithm ([22]). Indeed, the Mac-Queen algorithm takes mini-batches of size 1, and estimates the population of the cell j at the t-th iteration via $\sum_{\ell=1}^{t} \hat{p}_j^{(\ell)}$ instead of $t\hat{p}_j^{(t)}$, where $\hat{p}_j^{(t)} = \bar{X}_{B_t}\left(W_j(\mathbf{c}^{(t)})\right)$. These modifications are motivated by Theorem 5, that guarantees near-optimality of the output of Algorithm 2, provided that the mini-batches are large enough.

2.3. Theoretical guarantees

This section exposes theoretical guarantees for the two algorithms introduced in Section 2.2. Note that these guarantees are stated on the excess distortion $R(\mathbf{c}_T) - R^*$, where $\mathbf{c}_T$ is the output of the considered algorithm. In fact, the

same bounds hold also for $\|\mathbf{c}_T - \mathbf{c}^*\|^2$, up to the $M(\mathbb{E}(X))$ factor. A special interest will be paid to the sample-size dependency of the excess distortion. From this standpoint, a first negative result may be derived from the quantization framework. Indeed, from [20, Proposition 7], we may deduce that, for any empirically designed codebook $\hat{\mathbf{c}}$,

$$\inf_{\{X|\mathbb{E}(X) \text{ has a } r_0\text{-margin}\}} \mathbb{E}(R(\hat{\mathbf{c}}) - R^*) \geq c_0 M(E(X)) R^2 \frac{k^{1-\frac{2}{d}}}{n}. \tag{1}$$

In fact, this bounds holds in the special case where X satisfies the additional assumption $X = \delta_{X^{(1)}}$ a.s., pertaining to the vector quantization case. Thus it holds in the general case. This small result ensures that the sample-size dependency of the minimax excess distortion over the class of distribution of discrete measures whose mean measure satisfies a margin condition with radius r_0 is of order $1/n$ or greater.

A first upper bound on this minimax excess distortion may be derived from the following Theorem 4, that investigates the performance of the output of Algorithm 1.

Theorem 4. *Assume that $\mathbb{E}(X)$ satisfies a margin condition with radius r_0, and denote by $R_0 = \frac{Br_0}{16\sqrt{2}R}$, $\kappa_0 = \frac{R_0}{R}$. Choose $T = \lceil \frac{\log(n)}{\log(4/3)} \rceil$, and let $\mathbf{c}^{(T)}$ denote the output of Algorithm 1. If $\mathbf{c}^{(0)} \in \mathcal{B}(\mathcal{C}_{opt}, R_0)$, then, for n large enough, with probability $1 - 24e^{-c_1 n p_{min}^2 \kappa_0^2 / M^2} - e^{-x}$, where c_1 is a constant, we have*

$$R(\mathbf{c}^{(T)}) - R^* \leq M(\mathbb{E}(X)) \left(\frac{B^2 r_0^2}{512 R^2 n} + C \frac{M^2 R^2 k^2 d \log(k)}{n p_{min}^2} (1 + x) \right),$$

for all $x > 0$, where C is a constant.

Combined with (1), Theorem 4 ensures that Algorithm 1 reaches the minimax precision rate in terms of excess distortion after $O(\log(n))$ iterations, provided that the initialization is good enough. In the standard quantization case, Theorem 4 might be compared with [18, Theorem 3.1] for instance. In this case, the dependency on the dimension d provided by Theorem 4 is sub-optimal. Slightly anticipating, dimension-free bounds in the mean-measure quantization case exist, for instance by considering the output of Algorithm 2.

In practice, Theorem 4 guarantees that choosing $T = 2\log(n)$ and repeating several Loyd algorithms starting from different initializations provides an optimal quantization scheme. Note that combining [20, Theorem 3] or [30] and a deviation inequality for distortions such as in [18] gives an alternative proof of the optimality of Lloyd type schemes, in the sample points case where $X_i = \delta_{X_i^{(1)}}$. Theorem 4 provides in addition an upper bound on the number of iterations needed, as well as an extension of these results to the quantization of mean measure case. Its proof, that may be found in Section 5.1, relies on stochastic gradient techniques in the convex and non-smooth case. Bounds for the single-pass Algorithm 2 might be stated the same way.

Theorem 5. *Assume that $\mathbb{E}(X)$ satisfies a margin condition with radius r_0, and denote by $R_0 = \frac{Br_0}{16\sqrt{2}R}$, $\kappa_0 = R_0/R$. If $(B_1, \dots, B_T)$ are equally sized mini-batches of length $ckM^2 \log(n)/(\kappa_0 p_{min})^2$, where c is a positive constant, and $\mathbf{c}^{(T)}$ denotes the output of Algorithm 2, then, provided that $\mathbf{c}^{(0)} \in \mathcal{B}(\mathcal{C}_{opt}, R_0)$, we have*

$$\mathbb{E}\left(R(\mathbf{c}^{(T)}) - R^*\right) \leq M(\mathbb{E}(X))\left(Ck^2M^3R^2\frac{\log(n)}{n\kappa_0^2 p_{min}^3}\right).$$

A proof of Theorem 5 is given in Section 5.3. Theorem 5 entails that the resulting codebook of Algorithm 2 has an optimal distortion, up to a $\log(n)$ factor and provided that a good enough initialization is chosen. As for Algorithm 1, in practice, several initializations may be tried and the codebook with the best empirical distortion is chosen. Note that Theorem 5 provides a bound on the expectation of the distortion. Crude deviation bounds can be obtained using for instance a bounded difference inequality (see, e.g., [5, Theorem 6.2]). In the point sample case, more refined bounds can be obtained, using for instance [18, Theorem 4.1, Proposition 4.1]. To investigate whether these kind of bounds still hold in the measure sample case is beyond the scope of the paper. Note also that the bound on the excess distortion provided by Theorem 5 does not depend on the dimension d. This is also the case in [18, Theorem 3.1], where a dimension-free theoretical bound on the excess distortion of an empirical risk minimizer is stated in the sample points case. Interestingly, this bound also has the correct dependency in n, namely $1/n$. According to Theorem 4 and 5, providing a quantization scheme that provably achieves a dimension-free excess distortion of order $1/n$ in the sample measure case remains an open question.

3. Clustering of measures based on the quantized mean measure

3.1. Vectorization of measures

This Section introduces a vectorization method for measures, based on the quantization of the mean measure, that preserves separation between clusters if any. The intuition is the following: for a codebook $\mathbf{c} = (c_1, \dots, c_k)$ and a scale r, we may represent a discrete measure X via the vector of weights $(X(\mathcal{B}(c_1, r)), \dots, X(\mathcal{B}(c_k, r)))$ that encodes the mass that X spreads around every pole c_j. Now, if $X^{(1)}$ and $X^{(2)}$ are measures such that $|X^{(1)}(B(c_{j_0}, r)) - X^{(2)}(B(c_{j_0}, r))|$ is large, for some j_0, then the representations of $X^{(1)}$ and $X^{(2)}$ will be well separated. In practice, convolution with kernels is often preferred to local masses (see, e.g., [28]). To ease computation, we will restrict ourselves to the following class of kernel functions.

Definition 6. *For $(p, \delta) \in \mathbb{N}^* \times [0, 1/2]$, a function $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ is called a (p, δ)-kernel function if*

- $\|\psi\|_\infty \leq 1$,
- $\sup_{|u| \leq 1/p} \psi(u) \geq 1 - \delta$,

- $\sup_{|u|>2p} \psi(u) \leq \delta$,
- ψ *is* 1*-Lipschitz.*

Note that a (p, δ)-kernel is also a (q, δ)-kernel, for $q > p$. This definition of a kernel function encompasses widely used kernels, such as Gaussian or Laplace kernels. In particular, the function $\psi(u) = exp(-u)$ that is used in [28] is a $(p, 1/p)$-kernel for $p \in \mathbb{N}^*$. The 1-Lispchitz requirement is not necessary to prove that the representations of two separated measures will be well-separated. However, it is a key assumption to prove that the representations of two measures from the same cluster will remain close in $\mathbb{R}^k$. From a theoretical viewpoint, the more convenient kernel is $\psi_0 : x \mapsto (1-((x-1)\vee 0))\vee 0$, which is a $(1, 0)$-kernel, thus a $(p, 0)$-kernel for all $p \in \mathbb{N}^*$.

From now on we assume that the kernel ψ is fixed, and, for a k-points codebook $\mathbf{c}$ and scale factor σ, consider the vectorization

$$v_{\mathbf{c},\sigma} : \begin{cases} \mathcal{M}(R, M) & \to & [0, M]^k \\ X & \mapsto & (X(du) \bullet \psi(\|u - c_1\|/\sigma), \ldots, X(du) \bullet \psi(\|u - c_k\|/\sigma)) \end{cases} \tag{2}$$

Note that the dimension of the vectorization depends on the cardinality of the codebook $\mathbf{c}$. To guarantee that such a vectorization is appropriate for a classification purpose is the aim of the following section.

3.2. Discrimination and clustering based on the mean measure

In this section we investigate under which conditions the vectorization exposed in the above section provides a representation that is provably suitable for clustering. To this aim, for the sample $X_1, \ldots, X_n$, we introduce $(Z_1, \ldots, Z_n) \in [\![1, L]\!]^n$ the vector of (hidden) label variables. As well, we let $M_1, \ldots, M_L$ be such that, if $Z_i = \ell$, $X_i \in \mathcal{M}(R, M_\ell)$, and denote by $M = \max_{\ell \leq L} M_\ell$. For a given codebook $\mathbf{c}$, we introduce the following definition of (p, r, Δ)-scattering to quantify how well $\mathbf{c}$ will allow to separate clusters via the related vectorization.

Definition 7. *Let* $(p, r, \Delta) \in (\mathbb{N}^* \times \mathbb{R}^+ \times \mathbb{R}^+)$. *A codebook* $\mathbf{c} \in \mathcal{B}(0, M)^k$ *is said to* (p, r, Δ) ***-shatter*** $X_1, \ldots, X_n$ *if, for any* $i_1, i_2 \in [\![1, n]\!]$ *such that* $Z_{i_1} \neq Z_{i_2}$, *there exists* $j_{i_1,i_2} \in [\![1, k]\!]$ *such that*

$$X_{i_1}(\mathcal{B}(c_{j_{i_1,i_2}}, r/p)) \geq X_{i_2}(\mathcal{B}(c_{j_{i_1,i_2}}, 4pr)) + \Delta,$$

or

$$X_{i_2}(\mathcal{B}(c_{j_{i_1,i_2}}, r/p)) \geq X_{i_1}(\mathcal{B}(c_{j_{i_1,i_2}}, 4pr)) + \Delta.$$

In a nutshell, the codebook $\mathbf{c}$ shatters the sample if two different measures from two different clusters have different masses around one of the codepoint of $\mathbf{c}$, at scale r. Note that, for any i, j, $X_i(\mathcal{B}(c_j, r/p)) \geq X_i(\{c_j\})$, so that a stronger definition of shattering in terms of $X_i(\{c_j\})$'s might be stated, in the particular case where $X_i(\{c_j\}) > 0$. The following Proposition ensures that a codebook which shatters the sample yields a vectorization into separated clusters, provided the kernel decreases fast enough.

Proposition 8. *Assume that* $\mathbf{c} \in \mathcal{B}(0,R)^k$ *shatters* $X_1, \dots, X_n$*, with parameters* (p, r, Δ)*. Then, if* Ψ *is a* (p,δ)*-kernel, with* $\delta \leq \frac{\Delta}{4M}$*, we have, for all* i_1*,* i_2 $\in [\![1,n]\!]$*,*

$$Z_{i_1} \neq Z_{i_2} \quad \Rightarrow \quad \|v_{\mathbf{c},\sigma}(X_{i_1}) - v_{\mathbf{c},\sigma}(X_{i_2})\|_\infty \geq \frac{\Delta}{2},$$

for $\sigma \in [r, 2r]$*.*

A proof of Proposition 8 can be found in Section 6.1. This proposition shed some light on how $X_1, \dots, X_n$ has to be shattered with respect to the parameters of Ψ. Indeed, assume that $\Delta = 1$ (that is the case if the X_i's are integer-valued measures, such as count processes for instance). Then, to separate clusters, one has to choose δ small enough compared to $1/M$, and thus p large enough if Ψ is non-increasing. Hence, the vectorization will work roughly if the support points of two different counting processes are rp-separated, for some scale r. This scale r will then drive the choice of the bandwith σ. As shown in the following Section 4.2, this will be the case if the sample measures are persistence diagrams of well separated shapes. If the requirements of Proposition 8 are fulfilled, then a standard hierarchical clustering procedure such as Single Linkage with L_∞ distance will separate the clusters for the scales smaller than $\Delta/2$.

Now, to achieve a perfect clustering of the sample based on our vectorization scheme, we have to ensure that measures from the same cluster are not too far in terms of Wasserstein distance, implying in particular that they have the same total mass. This motivates the following definition.

Definition 9. *The sample of measures* $X_1, \dots X_n$ *is called* w*-concentrated if, for all* i_1*,* i_2 *in* $[\![1,n]\!]$ *such that* $Z_{i_1} = Z_{i_2}$*,*

- $X_{i_1}(\mathbb{R}^d) = X_{i_2}(\mathbb{R}^d)$,
- $W_1(X_{i_1}, X_{i_2}) \leq w$.

It now falls under the intuition that well-concentrated and shattered sample measures are likely to be represented in $\mathbb{R}^k$ by well-clusterable points. A precise statement is given by the following Proposition 10.

Proposition 10. *Assume that* $X_1, \dots, X_n$ *is* w*-concentrated. If* Ψ *is* 1*-Lipschitz, then, for all* $\mathbf{c} \in \mathcal{B}(0,R)^k$ *and* $\sigma > 0$*, for all* i_1*,* i_2 *in* $[\![1,n]\!]$ *such that* $Z_{i_1} = Z_{i_2}$*,*

$$\|v_{\mathbf{c},\sigma}(X_{i_1}) - v_{\mathbf{c},\sigma}(X_{i_2})\|_\infty \leq \frac{w}{\sigma}.$$

Therefore, if $X_1, \dots, X_n$ *is* (p, r, Δ)*-shattered by* $\mathbf{c}$*, and* $(r\Delta/4)$*-concentrated, then, for any* (p,δ)*-kernel satisfying* $\delta \leq \frac{\Delta}{4M}$*, we have*

$$\begin{aligned} Z_{i_1} = Z_{i_2} \quad &\Rightarrow \quad \|v_{\mathbf{c},\sigma}(X_{i_1}) - v_{\mathbf{c},\sigma}(X_{i_2})\|_\infty \leq \tfrac{\Delta}{4}, \\ Z_{i_1} \neq Z_{i_2} \quad &\Rightarrow \quad \|v_{\mathbf{c},\sigma}(X_{i_1}) - v_{\mathbf{c},\sigma}(X_{i_2})\|_\infty \geq \tfrac{\Delta}{2}, \end{aligned}$$

for $\sigma \in [r, 2r]$*.*

A proof of Proposition 10 is given in Section 6.2. An immediate consequence of Proposition 10 is that (p, r, Δ)-shattered and $r\Delta/4$-concentrated sample measures can be vectorized in $\mathbb{R}^k$ into a point cloud that is structured in L clusters. These clusters can be exactly recovered via Single Linkage clustering, with stopping parameter in $]\Delta/4, \Delta/2]$. In practice, tuning the parameter σ is crucial. Some heuristic is proposed in [28] in the special case of i.i.d persistence diagrams. An alternative calibration strategy is proposed in the following Section 4.2.

At last, from Propositions 8 and 10, if an optimal k-codebook of the mean measure shatters well the sample, then we can prove that the output of Algorithm 1 provides a relevant vectorization, with high probability. To properly define the mean measure in this case, we assume that the sample measures $X_1, \dots, X_n$ are drawn from a mixture model X. We let $Z \in [\![1, L]\!]$ denote a latent variable, with $\mathbb{P}(Z = \ell) = \pi_\ell$, and we assume that

$$X \mid \{Z = \ell\} \sim X^{(\ell)},$$

where $X^{(\ell)} \in \mathcal{M}(R, M_\ell)$, or equivalently $X = X^{(Z)}$. We also denote by $\bar{M} = \sum_{\ell=1}^{L} \pi_\ell M_\ell$, so that $\mathbb{E}(X) \in \mathcal{M}(R, \bar{M})$. In this framework, provided that $\mathbf{c}^*$ shatters well $X_1, \dots, X_n$, so will $\hat{\mathbf{c}}_n$, where $\hat{\mathbf{c}}_n$ is built with Algorithm 1.

Corollary 11. *Assume that $\mathbb{E}(X)$ satisfies the assumption of Theorem 4, and that $\mathbf{c}^*$ provides a (p, r, Δ) shattering of $X_1, \dots, X_n$, with $p \geq 2$. Let $\hat{\mathbf{c}}_n$ denote the output of Algorithm 1. Then, with probability larger than $1 - \exp\left[-C\left(\frac{nr^2 p_{min}^2}{p^2 M^2 R^2 k^2 d \log(k)} - \frac{p_{min}^2 B^2 r_0^2}{M^2 R^4 k^2 d \log(k)}\right)\right]$, $\hat{\mathbf{c}}_n$ is a $(\lfloor \frac{p}{2} \rfloor, r, \Delta)$ shattering of $X_1, \dots, X_n$, where C is a constant.*

A proof of Corollary 11 is given in Section 6.3. To fully assess the relevance of our vectorization technique, it remains to prove that k-points optimal codebooks for the mean measure provide a shattering of the sample measure, with high probability. This kind of result implies more structural assumptions on the components of the mixture X. The following Section 4.1 investigates the case where the sample measures are in fact persistence diagrams from differents shapes. In this particular case, we can show that quantization of the mean diagram is a relevant strategy to extract shattering codebooks.

4. Application for persistence diagrams

4.1. Mean measure of persistence diagrams

In this section we investigate the properties of our mean-measure quantization scheme in a particular instance of i.i.d. measure observations. Indeed, we assume that we observe n i.i.d persistence diagrams D_i, that are thought of as discrete measures on the half-plane $H^+ = \{(b, d) \in \mathbb{R}^2 \mid 0 \leq b \leq d\}$. In other words, the observations consist in n discrete measures $D_i = \sum_{j=1}^{n_i} \mu_{i,j} \delta_{x_{i,j}}$, where $x_{i,j} \in H^+$ and $\mu_{i,j}$ are weights that can be tuned beforehand. We will show that, whenever the persistence diagrams are generated from different samplings of the same shape, the mean persistence diagram and its best k-points

approximation are relevant topological features. Then, in a mixture of shapes framework, we will show that the mean persistence diagram might be used to build a vectorization of the persistence diagrams that allows a provably correct classification. In this section, a compact d-dimensional submanifold M of $\mathbb{R}^D$ is given, with positive reach τ_M (see, e.g., [24]). The object of interest will be the thresholded persistence diagram generated via the distance to M, denoted by d_M (where the infinite connected component has been removed). Namely, if $D' = \sum_{x \in H^+} n(x)\delta_x$ is the persistence diagram of d_M ($n(x)$ denotes the multiplicity of x), we aim to recover

$$D = \sum_{\{(b,d) \in D | d-b \geq s\}} n(b,d)\delta_{(b,d)} := \sum_{j=1}^{k_0} n(m_j)\delta_{m_j},$$

where the m_j's satisfy $m_j^2 - m_j^1 \geq s$. In general, such a thresholded diagram might have an infinite number of points, that is $k_0 = +\infty$. Whenever M is a compact set of $\mathbb{R}^d$, the following lemma ensures that k_0 is finite.

Lemma 12. *Let M be a compact subset of $\mathbb{R}^d$. The persistence diagram of the distance function d_M is denoted by D and, for any $s > 0$, the truncated diagram consisting of the points $m = (m^1, m^2) \in D$ such that $m^2 - m^1 \geq s$ is finite.*

A proof of Lemma 12 is given in Section 8.1. Further, we let P denote a distribution on M that has a density $f(x)$ with respect to the Hausdorff measure on M, bounded from below by f_{min}. We generate the sample persistence diagrams as follows: for $i = 1, \ldots, n$, $\mathbb{Y}_N^i$ denote an i.i.d N-sample drawn from M. According to Lemma 13 below, the distance to $\mathbb{Y}_N^i$ is a provably good approximation of d_M.

Lemma 13. *[1, Lemma B.7] Let $M \subset \mathbb{R}^D$ be a d-dimensional submanifold with positive reach τ_M, and let $\mathbb{Y}_N = Y_1, \ldots, Y_N$ be an i.i.d. sample drawn from a distribution that has a density $f(x)$ with respect to the Hausdorff measure of M. Assume that for all $x \in M$, $0 < f_{min} \leq f(x)$, and let $h = \left(\frac{C_d k \log(N)}{f_{min} N}\right)^{\frac{1}{d}}$, where C_d is a constant depending on d. If $h \leq \tau_M/4$, then, with probability larger than $1 - \left(\frac{1}{N}\right)^{\frac{k}{d}}$, we have*

$$\|d_M - d_{\mathbb{Y}_N}\|_\infty \leq h.$$

For convenience, in what follows we assume that for $i \leq n$, $\|d_M - d_{\mathbb{Y}_N^i}\|_\infty \leq h$, where $h = \left(\frac{C_d \log(N)}{f_{min} N}\right)^{\frac{1}{d}}$, for a constant C_d depending only on d. That occurs with high probability provided N is large enough. Then, for every $i = 1, \ldots, n$, if D_i' is the persistence diagram of the sublevel sets of $d_{\mathbb{Y}_n^i}$, we let

$$X_i = \sum_{\{x_{i,j} \in D_i' | x_{i,j}^2 - x_{i,j}^1 \geq s-h\}} \delta_{x_{i,j}}.$$

Note that $s \geq h$ provided N is large enough. This amounts to threshold the points of the persistence diagram D_i' that are close to the diagonal. The following

Lemma 14 ensures that X_i and D are close enough, in terms of bottleneck distance.

Lemma 14. *[11] If X and Y are compact sets of $\mathbb{R}^D$, then*

$$d_B\left(D(d_X), D(d_Y)\right) \leq \|d_X - d_Y\|_\infty.$$

This stability result allows us to state a result on the expected persistence diagram $\mathbb{E}(X)$. We recall that the thresholded persistence diagram of d_M is $D = \sum_{j=1}^{k_0} n(m_j)\delta_{m_j}$, and we denote by $\mathbf{m} = (m_1, \ldots, m_{k_0})$.

Proposition 15. *Let $h = \left(\frac{C_d \log(N)}{f_{min} N}\right)^{\frac{1}{d}}$. Then, for N large enough, with probability larger than $1 - \left(\frac{1}{N}\right)^{\frac{3}{d}}$, we have*

$$\|\mathbf{m} - \mathbf{c}^*\|_\infty \leq 8\sqrt{M}h,$$

where $\mathbf{c}^*$ is a k_0 optimal codebook for $\mathbb{E}(X)$ and $M = \sum_{j=1}^{k_0} n(m_j)$.

The proof of Proposition 15 is given in Section 7.1. If h is chosen small enough, Proposition 15 ensures that quantizing the expected persistence diagram yields a k_0-points distribution on the half-plane that is provably close to the targeted persistence diagram. This is of particular interest in the following Section 4.2, where we show that the mean persistence diagram provides a relevant feature in a mixture of shapes framework.

4.2. Vectorization and clustering of persistence diagrams

From the mean persistence diagram exposed in Section 4.1, we can build an embedding from the space of persistence diagrams to a finite-dimensional Euclidean space, that we will prove suitable for shape classification. A case of interest for shape classification is when X is a mixture distribution whose each component is drawn from a shape, as in Section 4.1. To be more precise, we let $L \in \mathbb{N}^*$ denote the number of components, and for $\ell \leq L$, we let $S^{(\ell)}$ denote a compact d_ℓ-submanifold, and $D^{(\ell)}_{\geq s}$ the thresholded persistence diagram built from $d_{S^{(\ell)}}$ (where points that have persistence smaller than s are removed).

As well, we denote by $X^{(\ell)}$ denote the distribution of the thresholded persistence diagram built from the distance to N_ℓ points drawn on $S^{(\ell)}$, with threshold $s - h_\ell$. Given a latent variable Z on $[\![1, L]\!]$, with $\mathbb{P}(Z = \ell) = \pi_\ell$, the mixture distribution X of thresholded persistence diagrams is given by

$$X \mid \{Z = \ell\} \sim X^{(\ell)},$$

or equivalently $X = X^{(Z)}$. To make discrimination between shapes possible, we have to assume that their persistence diagrams differ by at least one point.

Definition 16. *The shapes $S^{(1)}, \dots, S^{(\ell)}$ are **discriminable at scale** $\mathbf{s}$ if for any $1 \leq \ell_1 < \ell_2 \leq L$ there exists $m_{\ell_1,\ell_2} \in H^+$ such that*

$$D_{\geq s}^{(\ell_1)}(\{m_{\ell_1,\ell_2}\}) \neq D_{\geq s}^{(\ell_2)}(\{m_{\ell_1,\ell_2}\}),$$

where the thresholded persistence diagrams are considered as measures.

Note that if m_{ℓ_1,ℓ_2} satisfies the discrimination condition stated above, then $m_{\ell_1,\ell_2} \in D_{\geq s}^{(\ell_1)}$ or $m_{\ell_1,\ell_2} \in D_{\geq s}^{(\ell_2)}$. To discriminate between shapes, we have to ensure that every m_{ℓ_1,ℓ_2} is represented via an optimal codebook. This is the aim of the following Proposition.

Proposition 17. *Let $h_\ell = \left(\frac{C_{d_\ell}(d_\ell^2+2)\log(N_\ell)}{f_{min,\ell} N_\ell}\right)^{1/d_\ell}$, and $h = \max_{\ell \leq L} h_\ell$. Moreover, let $M_\ell = D_{\geq s}^{(\ell)}(H^+)$, $\bar{M} = \sum_{\ell=1}^L \pi_\ell M_\ell$, and $\pi_{min} = \min_{\ell \leq L} \pi_\ell$.*

Assume that $S^{(1)}, \dots, S^{(L)}$ are discriminable at scale s, and let $m_1, \dots, m_{k_0}$ denote the discrimination points. Let $K_0(h)$ denote

$$\inf\{k \geq 0 \mid \exists t_1, \dots, t_k \quad \bigcup_{\ell=1}^{L} D_{\geq s}^{(\ell)} \setminus \{m_1, \dots, m_{k_0}\} \subset \bigcup_{s=1}^{K_0(h)} \mathcal{B}_\infty(t_s, h)\}.$$

Let $k \geq k_0 + K_0(h)$, and $(c_1^*, \dots, c_k^*)$ denote an optimal k-points quantizer of $\mathbb{E}(X)$. Then, provided that N_ℓ is large enough for all ℓ, we have

$$\forall j \in [\![1, k_0]\!] \quad \exists p \in [\![1, k]\!] \quad \|c_p^* - m_j\|_\infty \leq \frac{5\sqrt{\bar{M}}h}{\sqrt{\pi_{min}}}.$$

The proof of Proposition 17 is given in Section 7.2. If $\bar{D}_{\geq s}$ denotes the mean persistence diagram $\sum_{\ell=1}^L \pi_\ell D_{\geq s}^{(\ell)}$, and $\bar{D}_{\geq s}$ has K_0 points, then it is immediate that $k_0 + K_0(h) \leq K_0$. Moreover, we also have $k_0 \leq \frac{L(L+1)}{2}$. Proposition 17 ensures that the discrimination points are well enough approximated by optimal k-centers of the expected persistence diagram $\mathbb{E}(X)$, provided the shapes $S^{(\ell)}$ are well-enough sampled and k is large enough so that $\bar{D}_{\geq s}$ is well-covered by k balls with radius h. Note that this is always the case if we choose $k = K_0$, but also allows for smaller k's.

In turn, provided that the shapes $S^{(1)}, \dots, S^{(L)}$ are discriminable at scale s and that k is large enough, we can prove that an optimal k-points codebook $\mathbf{c}^*$ is a (p, r, Δ)-shattering of the sample, with high probability.

Proposition 18. *Assume that the requirements of Proposition 17 are satisfied. Let $\tilde{B} = \min_{i=1,\dots,k_0, j=1,\dots,K_0, j\neq i} \|m_i - m_j\|_\infty \wedge s$. Let $\kappa > 0$ be a small enough constant. Then, if N_ℓ is large enough for all $\ell \in [\![1, \ell]\!]$, $X_1, \dots, X_n$ is $(p, r, 1)$-shattered by $\mathbf{c}^*$, with probability larger than $1 - n \max_{\ell \leq L} N_\ell^{-\left(\frac{(\kappa\tilde{B})^{d_\ell} f_{min,\ell} N_\ell}{C_\ell d_\ell \log(N_\ell)}\right)}$, provided that*

- $\frac{r}{p} \geq 2\kappa\tilde{B}$

- $4rp \leq \left(\frac{1}{2} - \kappa\right) \tilde{B}$.

Moreover, on this probability event, $X_1, \ldots, X_n$ is $2M\kappa\tilde{B}$-concentrated.

A proof of Proposition 18 is given in Section 7.2. In turn, Proposition 18 can be combined with Proposition 10 and Corollary 11 to provide guarantees on the output of Algorithm 1 combined with a suitable kernel. We choose to give results for the theoretical kernel $\psi_0 : x \mapsto (1 - ((x-1) \vee 0)) \vee 0$, and for the kernel used in [28], $\psi_{AT}(x) = \exp(-x)$.

Corollary 19. *Assume that the requirements of Proposition 18 are satisfied. For short, denote by v_i the vectorization of X_i based on the output of Algorithm 1. Then, with probability larger than* $1 - \exp\left[-C\left(\frac{nr^2p_{min}^2}{p^2M^2R^2k^2d\log(k)} - \frac{p_{min}^2B^2r_0^2}{M^2R^4k^2d\log(k)}\right)\right] - n\max_{\ell \leq L} N_\ell^{-\left(\frac{(\kappa\tilde{B})^{d_\ell} f_{min,\ell} N_\ell}{C_\ell d_\ell \log(N_\ell)}\right)}$, *where κ and C are small enough constants, we have*

$$\begin{aligned} Z_{i_1} = Z_{i_2} \quad &\Rightarrow \quad \|v_{i_1} - v_{i_2}\|_\infty \leq \tfrac{1}{4}, \\ Z_{i_1} \neq Z_{i_2} \quad &\Rightarrow \quad \|v_{i_1} - v_{i_2}\|_\infty \geq \tfrac{1}{2}, \end{aligned}$$

for $\sigma \in [r, 2r]$ and the following choices of p and r:

- *If $\Psi = \Psi_{AT}$, $p_{AT} = \lceil 4M \rceil$, and $r_{AT} = \frac{\tilde{B}}{32 p_{AT}}$.*
- *If $\Psi = \Psi_0$, $p_0 = 1$ and $r_0 = \frac{\tilde{B}}{32}$.*

A proof of Corollary 19 is given in Section 7.3. Corollary 19 can be turned into probability bounds on the exactness of the output of hierarchical clustering schemes applied to the sample points. For instance, on the probability event described by Corollary 19, Single Linkage with norm $\|.\|_\infty$ will provide an exact clustering. The probability bound in Corollary 19 shed some light on the quality of sampling of each shape that is required to achieve a perfect classification: roughly, for N_ℓ in $\Omega(\log(n))$, the probability of misclassification can be controlled. Note that though the key parameter $\tilde{B}$ is not known, in practice it can be scaled as several times the minimum distance between two points of a diagram.

References

[1] Aamari, E. and Levrard, C. (2019). Nonasymptotic rates for manifold, tangent space and curvature estimation. *Ann. Statist.* **47** 177–204. MR3909931

[2] Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F. and Ziegelmeier, L. (2017). Persistence images: a stable vector representation of persistent homology. *J. Mach. Learn. Res.* **18** Paper No. 8, 35. MR3625712

[3] Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F. and

Ziegelmeier, L. (2017). Persistence images: a stable vector representation of persistent homology. *Journal of Machine Learning Research* **18**.

[4] Boissonnat, J.-D., Chazal, F. and Yvinec, M. (2018). *Geometric and Topological Inference* **57**. Cambridge University Press.

[5] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration inequalities*. Oxford University Press, Oxford A nonasymptotic theory of independence, With a foreword by Michel Ledoux. MR3185193

[6] Carrière, M., Chazal, F., Ike, Y., Lacombe, T., Royer, M. and Umeda, Y. (2019). PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures. *arXiv e-prints* arXiv:1904.09378.

[7] Chazal, F., Cohen-Steiner, D., Guibas, L. J., Mémoli, F. and Oudot, S. Y. (2009). Gromov-Hausdorff Stable Signatures for Shapes using Persistence. *Computer Graphics Forum* **28** 1393-1403.

[8] Chazal, F., Cohen-Steiner, D. and Lieutier, A. (2009). A sampling theory for compact sets in Euclidean space. *Discrete & Computational Geometry* **41** 461–479.

[9] Chazal, F., de Silva, V., Glisse, M. and Oudot, S. (2016). *The structure and stability of persistence modules*. Springer International Publishing.

[10] Chazal, F. and Divol, V. (2018). The density of expected persistence diagrams and its kernel based estimation. In *34th International Symposium on Computational Geometry. LIPIcs. Leibniz Int. Proc. Inform.* **99** Art. No. 26, 15. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3824270

[11] Cohen-Steiner, D., Edelsbrunner, H. and Harer, J. (2007). Stability of persistence diagrams. *Discrete Comput. Geom.* **37** 103–120. MR2279866

[12] Cuturi, M. and Doucet, A. (2013). Fast Computation of Wasserstein Barycenters. *arXiv e-prints* arXiv:1310.4375.

[13] Diggle, P. J. (1990). A Point Process Modelling Approach to Raised Incidence of a Rare Phenomenon in the Vicinity of a Prespecified Point. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **153** 349-362.

[14] Edelsbrunner, H., Letscher, D. and Zomorodian, A. (2002). Topological persistence and simplification. **28** 511–533. Discrete and computational geometry and graph drawing (Columbia, SC, 2001). MR1949898

[15] Fischer, A. (2010). Quantization and clustering with Bregman divergences. *J. Multivariate Anal.* **101** 2207–2221. MR2671211

[16] Hoan Tran, Q., Vo, V. T. and Hasegawa, Y. (2018). Scale-variant topological information for characterizing the structure of complex networks. *arXiv e-prints* arXiv:1811.03573.

[17] Jung, I.-S., Berges, M., Garrett, J. H. and Poczos, B. (2015). Exploration and evaluation of AR, MPCA and KL anomaly detection techniques to embankment dam piezometer data. *Advanced Engineering Informatics* **29** 902 - 917. Collective Intelligence Modeling, Analysis, and Synthesis for Innovative Engineering Decision Making Special Issue of the 1st International Conference on Civil and Building Engineering Informatics.

[18] Levrard, C. (2015). Nonasymptotic bounds for vector quantization in

Hilbert spaces. *Ann. Statist.* **43** 592–619.
[19] Levrard, C. (2015). Supplement to "Non-asymptotic bounds for vector quantization".
[20] Levrard, C. (2018). Quantization/Clustering: when and why does k-means work. *Journal de la Société Française de Statistiques* **159**.
[21] Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28** 129–137. MR651807
[22] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)* Vol. I: Statistics, pp. 281–297. Univ. California Press, Berkeley, Calif. MR0214227
[23] Mendelson, S. and Vershynin, R. (2003). Entropy and the combinatorial dimension. *Invent. Math.* **152** 37–55. MR1965359
[24] Niyogi, P., Smale, S. and Weinberger, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.* **39** 419–441. MR2383768
[25] Rabin, J., Peyré, G., Delon, J. and Bernot, M. (2012). Wasserstein Barycenter and Its Application to Texture Mixing. In *Scale Space and Variational Methods in Computer Vision* (A. M. Bruckstein, B. M. ter Haar Romeny, A. M. Bronstein and M. M. Bronstein, eds.) 435–446. Springer Berlin Heidelberg, Berlin, Heidelberg.
[26] Rakhlin, A., Shamir, O. and Sridharan, K. (2011). Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. *arXiv e-prints* arXiv:1109.5647.
[27] Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G. and Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution* **6** 366-379.
[28] Royer, M., Chazal, F., Ike, Y. and Umeda, Y. (2019). ATOL: Automatic Topologically-Oriented Learning. *arXiv e-prints* arXiv:1909.13472.
[29] Shirota, S., Gelfand, A. E. and Mateu, J. (2017). Analyzing Car Thefts and Recoveries with Connections to Modeling Origin-Destination Point Patterns. *arXiv e-prints* arXiv:1701.05863.
[30] Tang, C. and Monteleoni, C. (2016). On Lloyd's Algorithm: New Theoretical Insights for Clustering in Practice. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* (A. Gretton and C. C. Robert, eds.). *Proceedings of Machine Learning Research* **51** 1280–1289. PMLR, Cadiz, Spain.
[31] van der Vaart, A. and Wellner, J. A. (2009). A note on bounds for VC dimensions. In *High dimensional probability V: the Luminy volume. Inst. Math. Stat. (IMS) Collect.* **5** 103–107. Inst. Math. Statist., Beachwood, OH. MR2797943
[32] Zieliński, B., Lipiński, M., Juda, M., Zeppelzauer, M. and Dłotko, P. (2018). Persistence Bag-of-Words for Topological Data Analysis. *arXiv e-prints* arXiv:1812.09245.

5. Proofs for Section 2

5.1. Proof of Theorem 4

Throughout this section we assume that $\mathbb{E}(X)$ satisfies a margin condition with radius r_0, and that $\mathbb{P}(X \in \mathcal{M}_{N_{max}}(R,M)) = 1$. We adopt the following notation: for any $\mathbf{c} \in \mathcal{B}(0,R)^k$, we denote by $\hat{p}_j(\mathbf{c}) = \bar{X}_n(W_j(\mathbf{c}))$, as well as $p_j(\mathbf{c}) = \mathbb{E}(X)(W_j(\mathbf{c}))$. Moreover, we denote by $\hat{m}(\mathbf{c})$ (resp. $m(\mathbf{c})$) the codebooks satisfying

$$\hat{m}(\mathbf{c})_j = \frac{\bar{X}_n(du)\left(u\mathbb{1}_{W_j(\mathbf{c})}(u)\right)}{\hat{p}_j(\mathbf{c})},$$
$$m(\mathbf{c})_j = \frac{\mathbb{E}(X)(du)\left(u\mathbb{1}_{W_j(\mathbf{c})}(u)\right)}{p_j(\mathbf{c})},$$

if $\hat{p}_j(\mathbf{c}) > 0$ (resp. $p_j(\mathbf{c}) > 0$), and $\hat{m}(\mathbf{c})_j = 0$ (resp. $m(\mathbf{c})_j = 0$) if $\hat{p}_j(\mathbf{c}) = 0$ (resp. $p_j(\mathbf{c}) = 0$). The proof of Theorem 4 makes intensive use of the following lemmas. The first lemma gathers concentration results

Lemma 20. *With probability larger than $1 - 8e^{-x}$, for all $\mathbf{c} \in \mathcal{B}(0,R)^k$,*

$$\hat{p}_j(\mathbf{c}) \leq p_j(\mathbf{c}) + \sqrt{\frac{4Mc_0kd\log(k)\log(2nN_{max})}{n} + \frac{4Mx}{n}}\sqrt{p_j(\mathbf{c})}$$
$$\hat{p}_j(\mathbf{c}) \geq p_j(\mathbf{c}) - \frac{4Mc_0kd\log(k)\log(2nN_{max})}{n} - \frac{4Mx}{n} - \sqrt{\frac{4Mc_0kd\log(k)\log(2nN_{max})}{n} + \frac{4Mx}{n}}\sqrt{p_j(\mathbf{c})},$$

where c_0 is an absolute constant. Moreover, with probability larger than $1 - e^{-x}$, we have

$$\sup_{\mathbf{c}\in\mathcal{B}(0,R)^k} \left\| \left((\bar{X}_n - \mathbb{E}(X))(du) \bullet \left[(c_j - u)\mathbb{1}_{W_j(\mathbf{c})}(u)\right]\right)_{j=1,\dots,k} \right\| \leq \frac{CRM}{\sqrt{n}}\left(k\sqrt{d\log(k)} + \sqrt{x}\right),$$

where C is a constant.

The proof of Lemma 20 is given in Section 8.2, and is based on empirical processes theory. The second Lemma roughly ensures that the gradient of the distortion function is Lipschitzian around optimal codebooks.

Lemma 21. *Assume that $\mathbb{E}(X) \in \mathcal{M}(R,M)$ satisfies a margin condition with radius r_0, and denote by $R_0 = \frac{Br_0}{16\sqrt{2}R}$. Let $\mathbf{c}^* \in \mathcal{C}_{opt}$, and $\mathbf{c}$ such that $\|\mathbf{c} - \mathbf{c}^*\| \leq R_0$. Then*

- $\sum_{j=1}^k |p_j(\mathbf{c}) - p_j(\mathbf{c}^*)| \leq \frac{p_{min}}{64}$,
- $\sum_{j=1}^k \|\mathbb{E}(X)(du)((u-c_j)\mathbb{1}_{W_j(\mathbf{c})}(u)) - p_j(\mathbf{c}^*)(c_j^* - c_j)\| \leq \frac{p_{min}}{8\sqrt{2}}\|\mathbf{c} - \mathbf{c}^*\|$.

The proof of Lemma 21 follows from [19, Section A.3]. At last, Lemma 22 below ensures that every step of Algorithm 1 is, up to concentration terms, a contraction towards an optimal codebook. We recall here that $R_0 = \frac{Br_0}{16\sqrt{2}R}$, $\kappa_0 = \frac{R_0}{R}$.

Lemma 22. *Assume that* $\mathbf{c} \in \mathcal{B}(\mathbf{c}^*, R_0)$. *Then, with probability larger than* $1 - 8e^{-c_1 n p_{min}/M}$, *for* n *large enough, we have,*

$$\|\hat{m}(\mathbf{c}) - \mathbf{c}^*\|^2 \leq \frac{3}{4}\|\mathbf{c} - \mathbf{c}^*\|^2 + \frac{K}{p_{min}^2} D_n^2,$$

where $D_n = \sup_{\mathbf{c}\in\mathcal{B}(0,R)^k} \| \left((\bar{X}_n - \mathbb{E}(X)) \bullet \left[(c_j - x)\mathbb{1}_{W_j(\mathbf{c})}(x)\right]\right)_{j=1,\dots,k} \|$ *and* K *is a positive constant.*

Furthermore, with probabilty larger than $1 - 8e^{-c_1 n p_{min}/M} - e^{-c_1 n p_{min}^2 \kappa_0^2/M^2}$, *for all* $\mathbf{c} \in \mathcal{B}(\mathbf{c}^*, R_0)$, *the above inequality holds and* $\hat{m}(\mathbf{c}) \in \mathcal{B}(\mathbf{c}^*, R_0)$, *for* n *large enough.*

The proof of Lemma 22 is postponed to the following Section 5.2. We are now in position to prove Theorem 4.

Proof of Theorem 4. First note that Algorithm 1 is defined by $\mathbf{c}^{(t+1)} = \hat{m}(\mathbf{c}^{(t)})$. Equipped with Lemma 22, the proof of Theorem 4 is straightforward. We settle on the event on which these Lemmas hold, that has probability larger than $1 - 24e^{-cn\kappa_0^2 p_{min}^2/M^2}$, for c small enough. On this event, we have that

$$\begin{aligned}\|\mathbf{c}^{(t)} - \mathbf{c}^*\|^2 &\leq \left(\frac{3}{4}\right)^t \|\mathbf{c}^{(0)} - \mathbf{c}^*\|^2 + \left(\sum_{p=0}^{t-1}\left(\frac{3}{4}\right)^p\right)\frac{K}{p_{min}^2} D_n^2 \\ &\leq \left(\frac{3}{4}\right)^t \|\mathbf{c}^{(0)} - \mathbf{c}^*\|^2 + \frac{4K}{p_{min}^2} D_n^2.\end{aligned}$$

Then we might bound D_n according to Proposition 20, that leads to the results. □

5.2. Proof of Lemma 22

Proof of Lemma 22. Let $\mathbf{c} \in \mathcal{B}(\mathbf{c}^*, R_0)$. We decompose $\|\hat{m}(\mathbf{c}) - \mathbf{c}^*\|^2$ as follows.

$$\|\hat{m}(\mathbf{c}) - \mathbf{c}^*\|^2 = \|\mathbf{c} - \mathbf{c}^*\|^2 + 2\langle \hat{m}(\mathbf{c}) - \mathbf{c}, \mathbf{c} - \mathbf{c}^*\rangle + \|\hat{m}(\mathbf{c}) - \mathbf{c}\|^2. \tag{3}$$

Next, we bound the first term of (3).

$$\begin{aligned}2\langle \hat{m}(\mathbf{c}) - \mathbf{c}, \mathbf{c} - \mathbf{c}^*\rangle &= 2\sum_{j=1}^k \frac{1}{\hat{p}_j(\mathbf{c})}\left\langle \bar{X}_n(du)((u - c_j)\mathbb{1}_{W_j(\mathbf{c})}(u)), c_j - c_j^*\right\rangle \\ &\leq 2\sum_{j=1}^k \frac{1}{\hat{p}_j(\mathbf{c})}\left\langle \mathbb{E}(X)(du)((u - c_j)\mathbb{1}_{W_j(\mathbf{c})}(u)), c_j - c_j^*\right\rangle + 2D_n\sqrt{\sum_{j=1}^k \frac{\|c_j - c_j^*\|^2}{\hat{p}_j(\mathbf{c})^2}} \\ &\leq 2\sum_{j=1}^k \frac{1}{\hat{p}_j(\mathbf{c})}\left\langle p_j(\mathbf{c}^*)(c_j^* - c_j), c_j - c_j^*\right\rangle \\ &\quad + \frac{2p_{min}}{8\sqrt{2}}\|\mathbf{c} - \mathbf{c}^*\|\sqrt{\sum_{j=1}^k \frac{\|c_j - c_j^*\|^2}{\hat{p}_j(\mathbf{c})^2}} + 2D_n\sqrt{\sum_{j=1}^k \frac{\|c_j - c_j^*\|^2}{\hat{p}_j(\mathbf{c})^2}},\end{aligned}$$

where the last line follows from Lemma 21. Now, using Lemma 20 with $x = c_1 n p_{min}/M$, for c_1 a small enough absolute constant, entails that, with probability larger than $1 - 8e^{c_1 n p_{min}/M}$, for n large enough and every $\mathbf{c} \in \mathcal{B}(\mathbf{c}^*, R_0)$,

$$\hat{p}_j(\mathbf{c}) \geq \frac{63}{64} p_j(\mathbf{c}) - \frac{p_{min}}{64} \geq \frac{31}{32} p_{min}$$
$$\hat{p}_j(\mathbf{c}) \leq \frac{33}{32} p_j(\mathbf{c}^*),$$

according to Lemma 21. Therefore

$$2\langle \hat{m}(\mathbf{c}) - \mathbf{c}, \mathbf{c} - \mathbf{c}^* \rangle \leq -2 \sum_{j=1}^{k} \frac{p_j(\mathbf{c}^*)}{\hat{p}_j(\mathbf{c})} \|c_j - c_j^*\|^2 + \frac{32}{124\sqrt{2}} \|\mathbf{c} - \mathbf{c}^*\|^2 + K_1 \|\mathbf{c} - \mathbf{c}^*\|^2 + K_1^{-1} \frac{32^2}{31^2 p_{min}^2} D_n^2, \quad (4)$$

where $K_1 > 0$ is to be fixed later. Then, the second term of (3) may be bounded as follows.

$$\|\hat{m}(\mathbf{c}) - \mathbf{c}\|^2 = \sum_{j=1}^{k} \frac{\|\bar{X}_n(du)((u - c_j)\mathbb{1}_{W_j(\mathbf{c})}(u))\|^2}{\hat{p}_j(\mathbf{c})^2} = \sum_{j=1}^{k} \frac{\left\|p_j(\mathbf{c}^*)(c_j - c_j^*) + \Delta_j(\mathbf{c}) + \Delta_{n,j}(\mathbf{c})\right\|^2}{\hat{p}_j(\mathbf{c})^2},$$

where

$$\Delta_j(\mathbf{c}) = \mathbb{E}(X)(du)\left[(u - c_j)\mathbb{1}_{W_j(\mathbf{c})}(u) - (u - c_j^*)\mathbb{1}_{W_j(\mathbf{c}^*)}(u)\right],$$

so that $\sum_{j=1}^{k} \|\Delta_j(\mathbf{c})\| \leq \frac{p_{min}}{8\sqrt{2}} \|\mathbf{c} - \mathbf{c}^*\|$, according to Lemma 21, and

$$\Delta_{n,j}(\mathbf{c}) = (\bar{X}_n - \mathbb{E}(X))(du)((u - c_j)\mathbb{1}_{W_j(\mathbf{c})}(u)),$$

so that $\sum_{j=1}^{k} \|\Delta_{n,j}\|^2 \leq D_n^2$. Thus,

$$\begin{aligned}
\|\hat{m}(\mathbf{c}) - \mathbf{c}\|^2 &\leq (1 + K_2 + K_3) \sum_{j=1}^{k} \frac{p_j(\mathbf{c}^*)^2}{\hat{p}_j(\mathbf{c})^2} \|c_j - c_j^*\|^2 + \left(1 + K_2^{-1} + K_4\right) \sum_{j=1}^{k} \frac{\|\Delta_j(\mathbf{c})\|^2}{\hat{p}_j(\mathbf{c})^2} \\
&\quad + \left(1 + K_3^{-1} + K_4^{-1}\right) \sum_{j=1}^{k} \frac{\|\Delta_{n,j}(\mathbf{c})\|^2}{\hat{p}_j(\mathbf{c})^2} \\
&\leq (1 + K_2 + K_3) \sum_{j=1}^{k} \frac{p_j(\mathbf{c}^*)^2}{\hat{p}_j(\mathbf{c})^2} \|c_j - c_j^*\|^2 + \left(1 + K_2^{-1} + K_4\right) \frac{32^2}{31^2 \times 128} \|\mathbf{c} - \mathbf{c}^*\|^2 \\
&\quad + \left(1 + K_3^{-1} + K_4^{-1}\right) \frac{32^2}{31^2 p_{min}^2} D_n^2, \quad (5)
\end{aligned}$$

wherer K_2, K_3 and K_4 are positive constants to be fixed later. Combining (4) and (5) yields that

$$\begin{aligned}\|\hat{m}(\mathbf{c}) - \mathbf{c}^*\|^2 \leq \|\mathbf{c} - \mathbf{c}^*\|^2 \left(1 + K_1 + \frac{32}{124\sqrt{2}} + \frac{32^2}{31^2 \times 128}\left(1 + K_2^{-1} + K_4\right)\right) \\ - 2\sum_{j=1}^{k} \frac{p_j(\mathbf{c}^*)}{\hat{p}_j(\mathbf{c})}\|c_j - c_j^*\|^2 + (1 + K_2 + K_3)\sum_{j=1}^{k} \frac{p_j(\mathbf{c}^*)^2}{\hat{p}_j(\mathbf{c})^2}\|c_j - c_j^*\|^2 \\ + D_n^2 \frac{32^2}{31^2 p_{min}^2}\left(1 + K_1^{-1} + K_3^{-1} + K_4^{-1}\right).\end{aligned}$$

Taking $K_2 = \frac{1}{32}$ gives, through numerical computation,

$$\begin{aligned}\|\hat{m}(\mathbf{c}) - \mathbf{c}^*\|^2 &\leq \|\mathbf{c} - \mathbf{c}^*\|^2 \left(0.62 + K_1 + K_3\frac{32^2}{31^2} + K_4\frac{32^2}{31^2 \times 128}\right) \\ &\quad + D_n^2 \frac{32^2}{31^2 p_{min}^2}\left(1 + K_1^{-1} + K_3^{-1} + K_4^{-1}\right) \\ &\leq \frac{3}{4}\|\mathbf{c} - \mathbf{c}^*\|^2 + \frac{K}{p_{min}^2} D_n^2,\end{aligned}$$

for K_1, K_3 and K_4 small enough. Now, according to Lemma 20, it holds

$$\frac{K}{p_{min}^2} D_n^2 \leq \frac{R_0^2}{4}$$

with probability larger than $1 - e^{-c_1 n p_{min}^2 \kappa_0^2 / M^2}$, for some constant c_1 small enough. This gives the second assertion of Lemma 22. □

5.3. *Proof of Theorem 5*

The proof of Theorem 5 will make use of the following deviation bounds.

Lemma 23. *Let $\mathbf{c} \in \mathcal{B}(0, R)^k$. Then, with probability larger than $1 - 2ke^{-x}$, we have, for all $j = 1, \ldots, k$,*

$$|\hat{p}_j(\mathbf{c}) - p_j(\mathbf{c})| \leq \sqrt{\frac{2Mp_j(\mathbf{c})x}{n}} + \frac{Mx}{n}.$$

Moreover, with probability larger than $1 - e^{-x}$, we have,

$$\left\|(\bar{X}_n - \mathbb{E}(X))(du) \bullet \left((c_j - u)\mathbb{1}_{W_j(\mathbf{c})}(u)\right)_{j=1,\ldots,k}\right\| \leq \frac{4RM\sqrt{k}}{\sqrt{n}}\left(1 + \sqrt{x}\right).$$

A proof of Lemma 23 is given in Section 8.3. Equipped with Lemma 23, the proof of Theorem 5 follows the proof of [26, Lemma 1].

Proof of Theorem 5. Assume that $n \geq k$. According to Lemma 23, if $n_t = |B_t| = \frac{CM}{p_{min}} \log(n)$, for C large enough to be fixed later, then, taking $x = 4\log(2n)$ leads to, for all $t \leq T$, with probability larger than $1 - \frac{k}{n^3}$,

$$|\hat{p}_j^{(t)} - p_j^{(t)}| \leq \frac{p_{min} + \sqrt{p_j^{(t)} p_{min}}}{256}$$
$$\left\| (\bar{X}_n - \mathbb{E}(X))(du) \bullet \left((c_j^{(t)} - u) \mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u) \right)_j \right\| \leq 8R\sqrt{\frac{kp_{min}}{CM}},$$

where $\hat{p}_j^{(t)}$ denotes $\bar{X}_{B_t}(W_j(\mathbf{c}^{(t)}))$. Let A_T denote this probability event. First we prove that if $\mathbf{c}^{(0)} \in \mathcal{B}(\mathbf{c}^*, R_0)$, then, on A_T, for all $t \leq T$, $\mathbf{c}^{(t)} \in \mathcal{B}(\mathbf{c}^*, R_0)$. We proceed recursively, assuming that $\mathbf{c}^{(t)} \in \mathcal{B}(\mathbf{c}^*, R_0)$. Then, on A_T, applying Lemma 21 yields that $\frac{33}{32} p_j^* \geq \hat{p}_j^t \geq \frac{31}{32} p_j^*$. Denoting by $a_t = \|\mathbf{c}^{(t)} - \mathbf{c}^*\|^2$ and $g_{t+1} = \left(\frac{\bar{X}_{B_{t+1}}(du) \bullet (c_j^{(t)} - u) \mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u)}{\hat{p}_j^{t+1}} \right)_j$, the recursion equation entails that

$$a_{t+1} \leq a_t - \frac{2}{t+1} \langle g_{t+1}, \mathbf{c} - \mathbf{c}^* \rangle + \frac{1}{(t+1)^2} \|g_{t+1}\|^2 . \tag{6}$$

As in the proof of Theorem 4, denote by

$$\Delta_j^t = \mathbb{E}(X)(du) \bullet \left((u - c_j^{(t)}) \mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u) \right) - p_j^*(c_j^* - c_j^{(t)})$$
$$\Delta_{n,j}^{t+1} = (\bar{X}_{B_{t+1}} - \mathbb{E}(X)(du)) \bullet (u - c_j^{(t)}) \mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u)$$
$$D_n^{t+1} = \sqrt{\sum_{j=1}^k \|\Delta_{n,j}^{t+1}\|^2}.$$

We have that

$$-\frac{2}{t+1} \langle g_{t+1}, \mathbf{c} - \mathbf{c}^* \rangle \leq -\frac{2}{t+1} \sum_{j=1}^k \left(\frac{p_j^*}{\hat{p}_j^{t+1}} \|c_j^{(t)} - c_j^*\|^2 - \frac{\|\Delta_{j,n}^{t+1}\| \|c_j^{(t)} - c_j^*\|}{\hat{p}_j^{t+1}} - \frac{\|c_j^{(t)} - c_j^*\| \|\Delta_j^t\|}{\hat{p}_j^{t+1}} \right)$$
$$\leq -2\frac{32}{33(t+1)} \|\mathbf{c}^{(t)} - \mathbf{c}^*\|^2 + \frac{64}{31p_{min}(t+1)} \|\mathbf{c}^{(t)} - \mathbf{c}^*\| D_n^{t+1} + \frac{64}{8\sqrt{2} \times 31(t+1)} \|\mathbf{c}^{(t)} - \mathbf{c}^*\|^2$$
$$\leq \|\mathbf{c}^{(t)} - \mathbf{c}^*\|^2 \left(\frac{-64}{33(t+1)} + \frac{K_4}{t+1} + \frac{64}{8\sqrt{2} \times 31(t+1)} \right) + K_4^{-1} \left(\frac{32}{31p_{min}} D_n^{t+1} \right)^2,$$

according to Lemma 21, where K_4 denotes a constant. Next, the second term

in (6) may be bounded by

$$\begin{aligned}\|g_{t+1}\|^2 \leq \sum_{j=1}^{k} \frac{1}{(\hat{p}_j^{t+1})^2}(p_j^*)^2\|c_j^{(t)} - c_j^*\|^2\left(1+K_1+K_2\right) + \frac{p_{min}^2}{128\min_j(\hat{p}_j^{t+1})^2}\|\mathbf{c}^{(t)}-\mathbf{c}^*\|^2\left(1+K_2^{-1}+K_3\right)\\ + \frac{1}{\min_j(\hat{p}_j^{t+1})^2}\left(1+K_1^{-1}+K_3^{-1}\right)(D_n^{t+1})^2\\ \leq \frac{32^2}{31^2}\|\mathbf{c}^{(t)}-\mathbf{c}^*\|^2\left(1+K_1+K_2+\frac{1+K_2^{-1}+K_3}{128}\right) + (D_n^{t+1})^2\frac{32^2(1+K_1^{-1}+K_3^{-1})}{31^2 p_{min}^2},\end{aligned}$$

where K_1, K_2 and K_3 are constants to be fixed later. Combining pieces and using $t+1 \geq 1$ leads to

$$\begin{aligned}a_{t+1} \leq a_t + \frac{a_t}{t+1}\left(\frac{-64}{33} + \frac{64}{8\sqrt{2}\times 31} + K_4 + \frac{32^2}{31^2}\left(1+K_1+K_2+\frac{1+K_2^{-1}+K_3}{128}\right)\right)\\ + \left(D_n^{t+1}\right)^2\left(\frac{32^2}{31^2 p_{min}^2}K_4^{-1} + \frac{32^2(1+K_1^{-1}+K_3^{-1})}{31^2 p_{min}^2}\right).\end{aligned}$$

Choosing $K_2 = \frac{1}{32}$ entails that

$$\begin{aligned}\left(\frac{-64}{33} + \frac{64}{8\sqrt{2}\times 31} + K_4 + \frac{32^2}{31^2}\left(1+K_1+K_2+\frac{1+K_2^{-1}+K_3}{128}\right)\right)\\ \leq -0.38 + K_4 + \frac{32^2}{31^2}\left(K_1 + \frac{K_3}{128}\right),\end{aligned}$$

so that, for K_1, K_3 and K_4 small enough, we have

$$a_{t+1} \leq 0.8 a_t + \frac{K}{p_{min}^2}(D_n^{t+1})^2.$$

Now, if $n_t = c_0\frac{kM^2}{p_{min}^2\kappa_0^2}$, $n \geq k$, where c_0 is an absolute constant, Lemma 23 with $x = 4\log(2n)$ yields that

$$a_{t+1} \leq 0.8 a_t + 0.2 R_0^2 \leq R_0^2.$$

Next, if $\mathcal{F}_t$ denotes the sigma-algebra corresponding to the observations of the t first mini-batches $B_1, \ldots, B_t$, and E_t denotes the conditional expectation with respect to $\mathcal{F}_t$. We will show that

$$\mathbb{E}a_{t+1} \leq \left(1 - \frac{2-K_1}{t+1}\right)\mathbb{E}a_t + \frac{12kMR^2}{p_{min}(t+1)^2}, \tag{7}$$

where $K_1 < \frac{1}{5}$. Starting from (6), we may write

$$\mathbb{E}a_{t+1} = \mathbb{E}a_t - \frac{2}{t+1}\mathbb{E}\left\langle g_{t+1}, \mathbf{c}^{(t)} - \mathbf{c}^*\right\rangle + \frac{1}{(t+1)^2}\mathbb{E}\|g_{t+1}\|^2.$$

Next, since on A_T it holds that $\left\langle \bar{X}_{B_{t+1}}(du)(c_j^{(t)}-u)\mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u), c_j^{(t)}-c_j^*\right\rangle \geq 0$, we may write

$$\begin{aligned}\mathbb{E}\left\langle -g_{t+1}, \mathbf{c}^{(t)}-\mathbf{c}^*\right\rangle &= \mathbb{E}\left(\left\langle -g_{t+1}, \mathbf{c}^{(t)}-\mathbf{c}^*\right\rangle \mathbb{1}_{A_T}\right) + R_1 \\ &\leq \mathbb{E}\sum_{j=1}^{k}\frac{32}{33p_j^*}\left\langle \bar{X}_{B_{t+1}}(u-c_j^t)\mathbb{1}_{W_j(\mathbf{c}^t)}(u), c_j^t-c_j^*\right\rangle \mathbb{1}_{A_T} + R_1 \\ &\leq \mathbb{E}\sum_{j=1}^{k}\frac{32}{33p_j^*}\left\langle \bar{X}_{B_{t+1}}(u-c_j^t)\mathbb{1}_{W_j(\mathbf{c}^t)}(u), c_j^t-c_j^*\right\rangle + R_1 + R_2,\end{aligned}$$

where, for $i=1,2$, $R_i \leq \frac{4kR^2M}{p_{min}}\mathbb{P}(A_T^c) \leq \frac{4k^2R^2M}{n^3 p_{min}} \leq \frac{4kMR^2}{p_{min}(t+1)^2}$. Next, we have that

$$\begin{aligned}E_t\sum_{j=1}^{k}\frac{32}{33p_j^*}&\left\langle \bar{X}_{B_{t+1}}(u-c_j^{(t)})\mathbb{1}_{W_j(\mathbf{c}^t)}(u), c_j^{(t)}-c_j^*\right\rangle \\ &\leq \sum_{j=1}^{k}\frac{32}{33p_j^*}E_t\left[\left\langle \bar{X}_{B_{t+1}}(u-c_j^t)\mathbb{1}_{W_j(\mathbf{c}^t)}(u), c_j^t-c_j^*\right\rangle\right] \\ &\leq -\frac{32}{33}a_t + \left(\sum_{j=1}^{k}\frac{32}{33p_j^*}\left\langle \Delta_j^{t+1}, c_j^t-c_j^*\right\rangle\right) \\ &\leq \left(-\frac{32}{33}\left(1-\frac{1}{8\sqrt{2}}\right)a_t\right),\end{aligned}$$

according to Lemma 21. Since $\|g_{t+1}\|^2 \leq 4kR^2$ and $p_{min} \leq \frac{M}{k}$, we immediatly get

$$\mathbb{E}a_{t+1} \leq \left(1-\frac{2-K_1}{t+1}\right)\mathbb{E}a_t + \frac{12kMR^2}{p_{min}(t+1)^2},$$

with $K_1 \leq 0.5$. Equipped with (7), we can prove Theorem 5, the same way as in the proof of [26, Lemma 1]. Namely, we prove recursively that

$$\mathbb{E}a_t \leq \frac{24kMR^2}{p_{min}t}.$$

Denote by $G = \frac{12kMR^2}{p_{min}}$. The case $t=1$ is obvious. Next, assuming that $\mathbb{E}a_t \leq \frac{2G}{t}$ and using (7) we may write

$$\begin{aligned}\mathbb{E}a_{t+1} &\leq \left(1-\frac{2}{t+1}\right)\mathbb{E}a_t + \frac{K_1}{t+1}\mathbb{E}a_t + \frac{G}{(t+1)^2} \\ &\leq \frac{G}{t(t+1)}\left[2t+2K_1-1\right].\end{aligned}$$

Since $K_1 \leq \frac{1}{2}$, we get that $\mathbb{E}a_{t+1} \leq 2G/(t+1)$. □

6. Proofs for Section 3

6.1. Proof of Proposition 8

Assume that $X_1, \dots, X_n$ is (p, r, Δ)-shattered by $\mathbf{c}$, let i_1, i_2 in $[\![1, n]\!]$ be such that $Z_{i_1} \neq Z_{i_2}$, and without loss of generality assume that

$$X_{i_1}(\mathcal{B}(c_1, r/p)) \geq \max_{u \in \mathcal{B}(c_1, r/p)} X_{i_2}(\mathcal{B}(u, 4rp)) + \Delta.$$

Let Ψ be a (p, δ)-kernel and $\sigma \in [r, 2r]$. We have

$$\begin{aligned} X_{i_1}(du) \bullet \Psi((u - c_1)/\sigma) &\geq X_{i_1}(du) \bullet \left[\Psi(\|u - c_1\|/\sigma) \mathbb{1}_{\mathcal{B}(c_1, r/p)}(u)\right] \\ &\geq (1 - \delta) X_{i_1}\left(\mathcal{B}(c_1, r/p)\right) \\ &\geq X_{i_1}\left(\mathcal{B}(c_1, r/p)\right) - \delta M. \end{aligned}$$

On the other hand, we have that

$$\begin{aligned} X_{i_2}(du) \bullet \Psi(\|u - c_1\|/\sigma) \leq\ & X_{i_2}(du) \bullet \left[\Psi(\|u - c_1\|/\sigma) \mathbb{1}_{\mathcal{B}(c_1, 4pr)}\right] \\ & + X_{i_2}(du) \bullet \left[\Psi(\|u - c_1\|/\sigma) \mathbb{1}_{(\mathcal{B}(c_1, 4pr))^c}\right] \\ &\leq X_{i_2}(\mathcal{B}(c_1, 4pr)) + \delta X_{i_2}((\mathcal{B}(c_1, 4pr))^c) \\ &\leq X_{i_1}\left(\mathcal{B}(c_1, r/p)\right) - \Delta + \delta M. \end{aligned}$$

We deduce that

$$\|v_{\mathbf{c},\sigma}(X_{i_1}) - v_{\mathbf{c},\sigma}(X_{i_2})\|_\infty \geq \Delta - 2\delta M \geq \frac{\Delta}{2},$$

whenever $\delta \leq \frac{\Delta}{4M}$.

6.2. Proof of Proposition 10

Let i_1, i_2 in $[\![1, n]\!]$ such that $Z_{i_1} = Z_{i_2}$. Let (Y_1, Y_2) be a random vector such that $Y_1 \sim X_{i_1}$, $Y_2 \sim X_{i_2}$, and $\mathbb{E}(\|Y_1 - Y_2\|) \leq w$. Let $c \in \mathcal{B}(0, R)$, we have

$$\begin{aligned} |X_{i_1}(du) \bullet \Psi(\|u - c\|/\sigma) - X_{i_2}(du) \bullet \Psi(\|u - c\|/\sigma)| & \\ &\leq \left|\mathbb{E}\left[\Psi(\|Y_1 - c\|/\sigma) - \Psi(\|Y_2 - c\|/\sigma)\right]\right| \\ &\leq \mathbb{E}\left(\frac{\|Y_1 - Y_2\|}{\sigma}\right) \\ &\leq \frac{w}{\sigma}, \end{aligned}$$

hence $\|v_{\mathbf{c},\sigma}(X_{i_1}) - v_{\mathbf{c},\sigma}(X_{i_2})\|_\infty \leq w/\sigma$. Now if $X_1 \dots, X_n$ is $r\Delta/4$-concentrated, and $\sigma \in [r, 2r]$, we have $\|v_{\mathbf{c},\sigma}(X_{i_1}) - v_{\mathbf{c},\sigma}(X_{i_2})\|_\infty \leq \frac{\Delta}{4}$.

6.3. Proof of Corollary 11

According to Theorem 4, for n large enough, with probability larger than $1 - \exp\left[-C\left(\frac{nr^2p_{min}^2}{p^2M^2R^2k^2d\log(k)} - \frac{p_{min}^2B^2r_0^2}{M^2R^4k^2d\log(k)}\right)\right]$, we have $\|\hat{\mathbf{c}}_n - \mathbf{c}^*\| \leq \frac{r}{p}$. Let i_1, $i_2 \in [\![1, n]\!]$ be such that $Z_{i_1} \neq Z_{i_2}$. Without loss of generality assume that

$$X_{i_1}\left(\mathcal{B}(c_1^*, r/p)\right) \geq X_{i_2}\left(\mathcal{B}(c_1^*, 4pr)\right) + \Delta.$$

Then $X_{i_1}\left(\mathcal{B}(\hat{c}_1, 2r/p)\right) \geq X_{i_1}\left(\mathcal{B}(\hat{c}_1, r/p)\right))$, and $X_{i_2}\left(\mathcal{B}(\hat{c}_1, 4(p/2)r)\right) \leq X_{i_2}\left(\mathcal{B}(c_1^*, 4pr)\right)$ entails that

$$X_{i_1}\left(\mathcal{B}(\hat{c}_1, 2r/p)\right) \geq X_{i_2}\left(\mathcal{B}(\hat{c}_1, 4(p/2)r)\right) + \Delta.$$

7. Proofs for Section 4

7.1. Proof of Proposition 15

Proof. Proof of Proposition 15 Let D_1' denote the persistence diagram build from the sublevel sets of $d_{\mathbb{Y}_N}$, where $\mathbb{Y}_N$ is an N-sample drawn on M (without the infinite connected component), and let R denote the diameter of M. Then, every point of D_1' is in $\mathcal{B}(0, R)$. For short denote by $\alpha_N = (\frac{1}{N})^{d+1/d}$, and we take N large enough so that $\alpha_N \leq \frac{h^2}{R^2} \wedge \frac{1}{2}$. For a positive t, we denote by $D_{\geq t} = \sum_{\{m \in D' | x^2 - x^1 \geq t\}} n(m)\delta_m$, where we recall that D' denotes the persistence diagram built from the sublevels sets of d_M. Since $D_{\geq \frac{s}{2}}$ is finite, there exists h_0 such that, for every $m \in D_{\geq s-h_0}$, $m \in D_{\geq s}$. At last, denote by $\tilde{B} = \min_{i \neq j} \|m_i - m_j\|_\infty$, where the m_j's are the points of $D_{\geq s}$, and choose n large enough so that $h \leq \frac{h_0}{2} \wedge \frac{\tilde{B}}{2}$.

For such an h, we have, with probability larger than $1 - \alpha_N$ so that $\|d_{\mathbb{Y}_N} - d_M\|_\infty \leq h$, for every $j \in [\![1, k_0]\!]$, $x_{i_1}^{(j)}, \ldots, x_{i_{n_j}}^{(j)} \in D_{1,\geq s-h} \cap \mathcal{B}_\infty(m_j, h)$, and $|D_{1,\geq s-h}| = M$. To bound $M(\mathbb{E}(D_{1,\geq s-h}))$, note that, with probability larger than $1-\alpha_N$, $M(D_{1,\geq s-h}) = M$, and with probability smaller than α_N, $M(D_{1,\geq s-h}) \leq N^d$, so that

$$|M(\mathbb{E}(D_{1,\geq s-h})) - M| \leq \alpha_N\left(M + N^d\right).$$

Next, we choose N large enough so that $|M(\mathbb{E}(D_{1,\geq s-h})) - M| \leq \frac{M}{2}$. Denoting by $\mathbf{m} = (m_1, \ldots, m_{k_0})$, we have

$$R(\mathbf{m}) \leq 2h^2M(1 - \alpha_N) + \alpha_N 4R^2 \times 3M/2 \leq 8Mh^2.$$

Now, if there exists j such that, for all $i \in [\![1, k_0]\!]$, $\|c_j - m_i\|_\infty > 8\sqrt{M}h$, then

$$R(\mathbf{c}) > (1 - \alpha_N)16Mh^2 \geq 8Mh^2 \geq R(\mathbf{m}).$$

□

7.2. *Proof of Proposition 17*

We let $\alpha_\ell = \left(\frac{1}{N_\ell}\right)^{d_\ell + \frac{2}{d_\ell}}$, and $A = \{\|d_{Y_{N_Z}} - d_{M_Z}\|_\infty > h_Z\}$, so that $\mathbb{P}(A \mid Z = \ell) \leq \alpha_\ell$. Also, let $m_{k_0+1}, \ldots, m_{k_0+K_0(h)}$ be such that $\bigcup_{\ell=1}^L D_{\geq s}^{(\ell)} \backslash \{m_1, \ldots, m_{k_0}\} \subset \bigcup_{s=1}^{K_0(h)} \mathcal{B}_\infty(m_{k_0+s}, h)$, and $\mathbf{m} = (m_1, \ldots, m_{k_0+K_0(h)})$. At last, we let $R = \max_{\ell \leq L} diam(S_\ell)$. For N_ℓ large enough so that $D_{\geq s-h_\ell}^{(\ell)} = D_{\geq s}^{(\ell)}$ and $s/2 > h_\ell$, we have

$$\begin{aligned}
R(\mathbf{m}) &= \mathbb{E}\left(\sum_{\ell=1} \mathbb{1}_{Z=\ell} X^{(\ell)}(du) \bullet \min_{j=1,\ldots,k_0+K_0(h)} \|u - m_j\|^2\right) \\
&= \mathbb{E}\left(\sum_{\ell=1} \mathbb{1}_{Z=\ell\cap A} X^{(\ell)}(du) \bullet \min_{j=1,\ldots,k_0+K_0(h)} \|u - m_j\|^2\right) \\
&\qquad + \mathbb{E}\left(\sum_{\ell=1} \mathbb{1}_{Z=\ell\cap A^c} X^{(\ell)}(du) \bullet \min_{j=1,\ldots,k_0+K_0(h)} \|u - m_j\|^2\right) \\
&\leq \mathbb{E}\left(\sum_{\ell=1} \mathbb{1}_{Z=\ell\cap A} 4R^2 N^{d_\ell}\right) + \mathbb{E}\left(\sum_{\ell=1} \mathbb{1}_{Z=\ell\cap A^c} M^{(\ell)} 2h_\ell^2\right) \\
&\leq 2h^2\bar{M} + 4R^2 \sum_{\ell=1}^L \pi_\ell \alpha_\ell N^{d_\ell}.
\end{aligned}$$

For N_ℓ large enough so that $\alpha_\ell N_\ell^{d_\ell} \leq \frac{\bar{M}h_\ell^2}{R^2}$, we have

$$R(\mathbf{m}) \leq 6h^2\bar{M}.$$

On the other hand, let $\mathbf{c}$ be a k-points codebook such that, for every $p \in [\![1, k]\!]$, $\|m_1 - c_p\|_\infty > 5\sqrt{\frac{\bar{M}}{\pi_{min}}}h$. Then we have

$$\begin{aligned}
R(\mathbf{c}) &\geq \mathbb{E}\left(\sum_{\ell=1}^L \mathbb{1}_{A^c\cap Z=\ell} X^{(\ell)}(\mathcal{B}_\infty(m_1, h)) \left(5\sqrt{\frac{\bar{M}}{\pi_{min}}} - 1\right)^2\right) h^2 \\
&\geq \mathbb{E}\left(\sum_{\ell=1}^L \mathbb{1}_{A^c\cap Z=\ell} n^{(\ell)}(m_1) \left(5\sqrt{\frac{\bar{M}}{\pi_{min}}} - 1\right)^2\right) h^2.
\end{aligned}$$

Now let ℓ_0 be such that $n^{(\ell_0)}(m_1) \geq 1$, and assume that the N_ℓ's are large enough so that $\alpha_\ell \leq \frac{1}{2}$. It holds

$$\begin{aligned}
R(\mathbf{c}) &\geq \left(5\sqrt{\frac{\bar{M}}{\pi_{min}}} - 1\right)^2 h^2 \pi_{\ell_0}(1 - \alpha_{\ell_0}) \\
&\geq 8\bar{M}h^2 \\
&> R(\mathbf{m}),
\end{aligned}$$

hence the result.

Proof of Proposition ??

We let $M = \max_{\ell \leq L} M_\ell$, h_0 be such that $\bar{D}_{\geq s-h_0} = \bar{D}_{\geq s}$. Let $\kappa \leq \frac{1}{16} \wedge \frac{h_0}{2\tilde{B}}$. Under the assumptions of Proposition 17, we choose N_ℓ, $\ell \leq L$ large enough so that

$$\frac{5\sqrt{M}h}{\sqrt{\pi_{min}}} \leq \kappa \tilde{B}.$$

Next, denote by $\alpha_\ell = N_\ell^{-\left(\frac{(\kappa\tilde{B})^{d_\ell} f_{min,\ell} N_\ell}{C_\ell d_\ell \log(N_\ell)}\right)}$. Then we have

$$\begin{aligned}
\mathbb{P}\left(\exists i \in [\![1,n]\!] \mid d_B(X_i, D^{Z_i}_{\geq s}) > \kappa\tilde{B}\right) &\leq \sum_{i=1}^{n} \mathbb{P}\left(d_B(X_i, D^{Z_i}) > \kappa\tilde{B}\right) \\
&\leq \sum_{i=1}^{n}\sum_{\ell=1}^{L} \pi_\ell \alpha_\ell \\
&\leq n \max_{\ell \leq L} \alpha_\ell.
\end{aligned}$$

For the remaining of the proof we assume that, for $i = 1, \ldots, n$, $d_B(X_i, D^{Z_i}_{\geq s}) \leq \kappa\tilde{B}$, that occurs with probability larger than $1 - n\max_{\ell \leq L} \alpha_\ell$. Let $i_1 \neq i_2$, and assume that $Z_{i_1} = Z_{i_2} = z$. Then $W_\infty(X_{i_1}, X_{i_2}) = d_B(X_{i_1}, X_{i_2}) \leq 2\kappa\tilde{B}$. Hence $W_1(X_{i_1}, X_{i_2}) \leq 2M\kappa\tilde{B}$.

Now assume that $Z_{i_1} \neq Z_{i_2}$, and without loss of generality $m_{Z_{i_1}, Z_{i_2}} = m_1$ with $D^{Z_{i_1}}_{\geq s}(\{m_1\}) \geq D^{Z_{i_2}}_{\geq s}(\{m_1\}) + 1$. Let (p, r) in $\mathbb{N}^* \times \mathbb{R}^+$ be such that $r/p \geq 2\kappa\tilde{B}$ and $4rp \leq \left(\frac{1}{2} - \kappa\right)\tilde{B}$. Since $\|c_1^* - m_1\|_\infty \leq \kappa\tilde{B}$ and $d_B(X_{i_1}, D^{Z_{i_1}}_{\geq s}) \leq \kappa\tilde{B} < h_0$, we get

$$X_{i_1}\left(\mathcal{B}(c_1^*, \frac{r}{p})\right) = D^{Z_{i_1}}_{\geq s}(\{m_1\}).$$

On the other hand, since $4rp \leq (\frac{1}{2} - \kappa)\tilde{B}$, we also have

$$X_{i_2}\left(\mathcal{B}(c_1^*, 4rp)\right) = D^{Z_{i_2}}_{\geq s}(\{m_1\}).$$

Thus $X_1, \ldots, X_n$ is $(p, r, 1)$-shattered by $\mathbf{c}^*$.

7.3. Proof of Corollary 19

In the case where $\Psi = \Psi_{AT}$, we have that Ψ_{AT} is a $(p, 1/p)$ kernel. The requirement $1/p \leq \frac{1}{4M}$ of Proposition 8 is thus satisfied for $p_{AT} = \lceil 4M \rceil$. On the other hand, choosing $r_{AT} = \frac{\tilde{B}}{32 p_{AT}}$ ensures that $8 r_{AT} p_{AT} \leq (1/2 - \kappa)\tilde{B}$ and $\frac{r_{AT}}{2p_{AT}} \geq 2\kappa\tilde{B}$, for κ small enough. Thus, the requirements of Proposition 18 are satisfied: $\mathbf{c}^*$ is a $(2p_{AT}, r, 1)$ shattering of $X_1, \ldots, X_n$). At last, using Corollary 11, we have that $\hat{\mathbf{c}}_n$ is a $(p_{AT}, r_{AT}, 1)$ shattering of $X_1, \ldots, X_n$, on the probability event described by Corollary 11. It remains to note that $2\kappa\tilde{B} \leq \frac{r_{AT}}{4}$

for κ small enough to conclude that $X_1, \ldots, X_n$ is $\frac{r_{AT}}{4}$-concentrated on the probability event described in Proposition 18. Thus Proposition 10 applies.

The case $\Psi = \Psi_0$ is simpler. Since Ψ_0 is a $(1,0)$-kernel, we obviously have that $0 \leq \frac{1}{2M}$, so that the requirement of Proposition 18 is satisfied. With $p_0 = 1$ and $r_0 = \frac{\tilde{B}}{16}$ we immediatly get that $r_0/(2p_0) \geq 2\kappa\tilde{B}$ and $8r_0p_0 \leq (1/2 - \kappa\tilde{B})$, for κ small enough, so that $\hat{\mathbf{c}}$ is a $(p_0, r_0, 1)$ shattering of $X_1, \ldots, X_n$. As well, $2M\kappa\tilde{B} \leq \frac{r_0}{4}$, for κ small enough. Thus Proposition 10 applies.

8. Technical proof

8.1. Proof of Lemma 12

The lemma follows from standard arguments in geometric inference and persistent homology theory.

First, the definition of generalized gradient of d_M - see [8] or [4] Section 9.2 - implies that the critical points of d_M are all contained in the convex hull of M. As a consequence, they are all contained in the sublevel set $d_M^{-1}([0, 2\text{diam}(M)])$. It follows from the Isotopy Lemma - [4] Theorem 9.5 - that all the sublevel sets $d_M^{-1}([0,t])$, $t > 2\text{diam}(M)$ have the same homology. As a consequence, no point in D has a larger coordinate than $2\text{diam}(M)$ and D is contained in $[0, 2\text{diam}(M)]^2$.

Since M is compact, the persistence module of the filtration defined by the sublevel sets of d_M is q-tame (Corollary 3.35 in [9]). Equivalently, this means that for any $b_0 < d_0$, the intersection of D with the quadrant $Q_{(b_0,d_0)} = \{(b,d) : b < b_0 \text{ and } d_0 < d\}$ is finite. Noting that the intersection of $[0, 2\text{diam}(M)]^2$ with the half-plane $\{(b,d) : d \geq b + s\}$ can be covered by a finite union of quadrants $Q_{(b,b+\frac{s}{2}}$ concludes the proof of the lemma.

8.2. Proof of Lemma 20

Let Z_1 denote the process

$$Z_1 = \sup_{\mathbf{c} \in \mathcal{B}(0,R)^k, j=1,\ldots,k} \left| \left(\frac{\bar{X}_n}{M} - \frac{\mathbb{E}(X)}{M} \right) \mathbb{1}_{W_j(\mathbf{c})} \right|.$$

Note that the VC dimensions of Voronoi cells in a k-points Voronoi diagram is at most $c_0 kd \log(k)$ ([31, Theorem 1.1]). We first use a symmetrization bound.

Lemma 24. *Let $\mathcal{F}$ denote a class of functions taking values in $[0,1]$, and $X_1, \ldots, X_n$, $X'_1, \ldots, X'_n$ i.i.d random variables drawn from P. Denote by P_n and P'_n the empirical distributions associated to the X_i's and X'_i's. If $nt^2 \geq 1$,*

then

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\frac{(P-P_n)f}{\sqrt{Pf}}\geq 2t\right)\leq 2\mathbb{P}\left(\sup_{f\in\mathcal{F}}\frac{(P'_n-P_n)f}{\sqrt{(P'_nf+P_nf)/2}}\geq t\right)$$
$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\frac{(P_n-P)f}{\sqrt{P_nf}}\geq 2t\right)\leq 2\mathbb{P}\left(\sup_{f\in\mathcal{F}}\frac{(P_n-P'_n)f}{\sqrt{(P'_nf+P_nf)/2}}\geq t\right).$$

For the sake of completeness a proof of Lemma 24 is given in Section 8.4. Next, introducing $\sigma_1,\dots,\sigma_n$ independent Rademacher variables, we get

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\frac{(P'_n-P_n)f}{\sqrt{(P'_nf+P_nf)/2}}\geq t\right)\leq\mathbb{P}\left(\sup_{f\in\mathcal{F}}\frac{\frac{1}{n}\sum_{i=1}^n\sigma_i(f(X_i)-f(X'_i))}{\sqrt{(P'_nf+P_nf)/2}}\geq t\right)$$
$$\leq\mathbb{E}_{X_1,\dots,X_n,X'_1,\dots,X'_n}\left(\mathbb{P}_\sigma\left(\sup_{f\in\mathcal{F}}\frac{\frac{1}{n}\sum_{i=1}^n\sigma_i(f(X_i)-f(X'_i))}{\sqrt{(P'_nf+P_nf)/2}}\geq t\right)\right).$$

For a set of functions $\mathcal{F}$ and elements $x_1,\dots,x_q\in\mathcal{M}(R)$ we denote by $S_{\mathcal{F}}(x_1,\dots,x_q)$ the cardinality of the set $\{(f(x_1),\dots,f(x_q))\mid f\in\mathcal{F}\}$. Let $\mathcal{F}_1$ denote the sets of functions $\{X\in\mathcal{M}(R,M)\mapsto X(W)/M\mid W=\bigcap_{j=1}^k H_j,\ H_j \text{ half-space}\}$. Since, for every $i\in[\![1,n]\!]$, $X_i=\sum_{j=1}^{n_i}\mu_{i,j}\delta_{x_j^{(i)}}$, we have

$$S_{\mathcal{F}_1}(X_1,\dots,X_n,X'_1,\dots,X'_n)\leq|\{(\mathbb{1}_W(x_j^{(i)}))_{i=1,\dots,2n,j=1,\dots,n_i}\mid W=\bigcap_{j=1}^k H_j,\ H_j \text{ half-space}\}|$$
$$\leq\left(2\left(\sum_{i=1}^n n_i+n'_i\right)\right)^{c_0kd\log(k)},$$

using [23, Theorem 1], and [31, Theorem 1] to bound the VC-dimension of the sets W's. On the other hand, for any $f\in\mathcal{F}_1$, it holds

$$\frac{\sum_{i=1}^n(f(X'_i)-f(X_i))^2}{\sum_{i=1}^n(f(X_i)+f(X'_i))}\leq 1.$$

Thus, combining Hoeffding's inequality and a plain union bound yields

$$\left(\mathbb{P}_\sigma\left(\sup_{f\in\mathcal{F}}\frac{\frac{1}{n}\sum_{i=1}^n\sigma_i(f(X_i)-f(X'_i))}{\sqrt{(P'_nf+P_nf)/2}}\geq t\right)\right)\leq\left(2\sum_{i=1}^n(n_i+n'_i)\right)^{c_0kd\log(k)}e^{-nt^2},$$

hence, since for $i=1,\dots,n$, $X_i\in\mathcal{M}_{N_{max}}(R,M)$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\frac{(P'_n-P_n)f}{\sqrt{(P'_nf+P_nf)/2}}\geq t\right)\leq(4nN_{max})^{c_0kd\log(k)}e^{-nt^2},$$

that proves the second inequality of Lemma 20. The first inequality of Lemma 20 derives the same way from the second inequality of Lemma 24.

We turn to the third inequality of Lemma 20. Let Z denote the process

$$Z = \sup_{\mathbf{c}\in\mathcal{B}(0,R)^k, \|\mathbf{t}\|\leq 1} \left\langle \left((\frac{\bar{X}_n}{M} - \frac{\mathbb{E}(X)}{M}) \bullet \left[(c_j - u)\mathbb{1}_{W_j(\mathbf{c})}(u) \right] \right)_{j=1,\dots,k}, \mathbf{t} \right\rangle,$$

and, for $j = 1, \dots, k$,

$$Z_j = \sup_{\mathbf{c}\in\mathcal{B}(0,R)^k, \|t_j\|\leq 1} \left\langle \frac{1}{M}(\bar{X}_n - \mathbb{E}(X)) \bullet \left[(c_j - u)\mathbb{1}_{W_j(\mathbf{c})}(u) \right], t_j \right\rangle,$$

so that $Z \leq \sqrt{\sum_{j=1}^k Z_j^2}$. According to the bounded differences inequality ([5, Theorem 6.2]), we have

$$\mathbb{P}\left(Z_j \geq \mathbb{E}(Z_j) + \sqrt{\frac{8R^2}{n}x} \right) \leq e^{-x}.$$

Using symmetrization we get

$$\mathbb{E}Z_j \leq \frac{2}{n}\mathbb{E}_{X_1,\dots,X_n}\mathbb{E}_\sigma \sup_{\mathbf{c}\in\mathcal{B}(0,R)^k, \|t\|\leq 1} \sum_{i=1}^n \sigma_i \left\langle \frac{X_i}{M} \bullet \left[(c_j - .)\mathbb{1}_{W_j(\mathbf{c})}(.) \right], t_j \right\rangle,$$

where $\sigma_1, \dots, \sigma_n$ are i.i.d Rademacher variables. Now assume that $X_1, \dots, X_n$ is fixed and $j = 1$. For a set $\mathcal{F}$ of real-valued functions we denote by $\mathcal{N}(\mathcal{F}, \varepsilon, \|.\|)$ its ε-covering number with respect to the norm $\|.\|$. Denoting by Γ_0, Γ_1 and Γ_2 the following sets

$$\begin{aligned}
\Gamma_0 &= \left\{ \gamma^{(0)}_{(\mathbf{c},t_1)} : X \mapsto \tfrac{X}{M} \bullet \left[\tfrac{\langle c_1 - ., t_1\rangle}{2R} \mathbb{1}_{W_1(\mathbf{c})}(.) \right] \mid \mathbf{c} \in \mathcal{B}(0,R)^k, t_1 \in \mathcal{B}(0,1) \right\}, \\
\Gamma_1 &= \left\{ \gamma^{(1)}_{(c_1,t_1)} : x \mapsto \tfrac{\langle c_1 - x, t_1\rangle}{2R} \mid c_1 \in \mathcal{B}(0,R), t_1 \in \mathcal{B}(0,1) \right\}, \\
\Gamma_2 &= \left\{ \gamma^{(2)}_{\mathbf{c}'} : x \mapsto \mathbb{1}_{W_1(\mathbf{c})}(x) \mid \mathbf{c} \in \mathcal{B}(0,R)^k \right\},
\end{aligned}$$

so that, for every $(\mathbf{c}, t_1), (\mathbf{c}', t_1') \in (\mathcal{B}(0,R)^k \times \mathcal{B}(0,1))^2$,

$$\gamma^{(0)}_{(\mathbf{c},t_1)}(X_i) - \gamma^{(0)}_{(\mathbf{c}',t_1')}(X_i) = \frac{X_i}{M} \bullet \left[\gamma^{(1)}_{(c_1,t_1)}(.)\gamma^{(2)}_{\mathbf{c}}(.) - \gamma^{(1)}_{(c_1',t_1')}(.)\gamma^{(2)}_{\mathbf{c}'}(.) \right].$$

Let $\varepsilon > 0$. If $\|\gamma^{(1)}_{(c_1,t_1)} - \gamma^{(1)}_{(c_1',t_1')}\|_\infty \leq \varepsilon$, we may write

$$(\gamma^{(0)}_{(\mathbf{c},t_1)}(X_i) - \gamma^{(0)}_{(\mathbf{c}',t_1')}(X_i))^2 \leq \left(\varepsilon + \frac{X_i}{M} \bullet \left[|\gamma^{(2)}_{\mathbf{c}} - \gamma^{(2)}_{\mathbf{c}'}| \right] \right)^2.$$

Thus,

$$\begin{aligned}
\|\gamma^{(0)}_{(\mathbf{c},t_1)} - \gamma^{(0)}_{(\mathbf{c}',t_1')}\|^2_{L_2(P_n)} &\leq 2\varepsilon^2 + \frac{2}{n}\sum_{j=1}^n \|\gamma^{(2)}_{\mathbf{c}} - \gamma^{(2)}_{\mathbf{c}'}\|^2_{L_2(X_i/M)} \\
&\leq 2\varepsilon^2 + 2\|\gamma^{(2)}_{\mathbf{c}} - \gamma^{(2)}_{\mathbf{c}'}\|^2_{L_2(\bar{X}_n/M)} \\
&\leq 2\varepsilon^2 + 2\|\gamma^{(2)}_{\mathbf{c}} - \gamma^{(2)}_{\mathbf{c}'}\|^2_{L_2(\bar{X}_n/M(\bar{X}_n))}.
\end{aligned}$$

We deduce

$$\mathcal{N}(\Gamma_0, \varepsilon, L_2(P_n)) \leq \mathcal{N}(\Gamma_1, \varepsilon/2, \|.\|_\infty) \times \mathcal{N}(\Gamma_2, \varepsilon/2, L_2(\bar{X}_n/M(\bar{X}_n))),$$

for every $\varepsilon > 0$. According to [23, Theorem 1], we may write

$$\mathcal{N}\left(\Gamma_1, \frac{\varepsilon}{2}, \|.\|_\infty\right) \leq \left(\frac{4}{\varepsilon}\right)^{K(d+1)}$$
$$\mathcal{N}\left(\Gamma_2, \frac{\varepsilon}{2}, L_2(\bar{X}_n/M(\bar{X}_n))\right) \leq \left(\frac{4}{\varepsilon}\right)^{c_0 K k d \log(k)},$$

where K is a constant and $\varepsilon < 2$. Thus, for every $\varepsilon < 2$,

$$\mathcal{N}\left(\Gamma_0, \varepsilon, L_2(P_n)\right) \leq \left(\frac{4}{\varepsilon}\right)^{Ckd\log(k)}.$$

Using Dudley's entropy integral (see, e.g., [5, Corollary 13.2]) yields, for $k \geq 2$,

$$\mathbb{E}_\sigma Z_j \leq CR\sqrt{\frac{kd\log(k)}{n}},$$

hence the result.

8.3. Proof of Lemma 23

The first bound of Lemma 23 follows from Bernstein's inequality. To prove the second inequality, we first bound the expectation as follows.

$$\begin{aligned}
&\mathbb{E}\left(\left\|(\bar{X}_n - \mathbb{E}(X))(du) \bullet \left((c_j - u)\mathbb{1}_{W_j(\mathbf{c})}(u)\right)_{j=1,\dots,k}\right\|\right) \\
&\qquad \leq \sqrt{\mathbb{E}\left\|(\bar{X}_n - \mathbb{E}(X))(du) \bullet \left((c_j - u)\mathbb{1}_{W_j(\mathbf{c})}(u)\right)_{j=1,\dots,k}\right\|^2} \\
&\qquad \leq \sqrt{\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left(\|(X_i - \mathbb{E}(X)) \bullet \left((c_j - u)\mathbb{1}_{W_j(\mathbf{c})}(u)\right)_{j=1,\dots,k}\|^2\right)} \\
&\qquad \leq \sqrt{\frac{(4RM)^2 k}{n}} = \frac{4RM\sqrt{k}}{\sqrt{n}}
\end{aligned}$$

A bounded difference inequality (see, e.g., [5, Theorem 6.2] entails that, with probability larger than $1 - e^{-x}$,

$$\left\|(\bar{X}_n - \mathbb{E}(X))(du) \bullet \left((c_j - u)\mathbb{1}_{W_j(\mathbf{c})}(u)\right)_{j=1,\dots,k}\right\| \leq \frac{4RM\sqrt{k}}{\sqrt{n}} + \sqrt{\frac{8kR^2M^2x}{n}},$$

hence the result.

8.4. Proof of Lemma 24