# Clustering of measures via mean measure quantization

**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 measures. Our goal is to build from data an embedding of these measures into a finite-dimensional Euclidean space, that allows for provably efficient clustering of the source measures.

The vectorization technique we propose relies on finding a compactly supported approximation of the mean measure generating process, that coincides with the intensity measure in the point process framework. To this aim we provide two algorithms that we prove almost minimax optimal.

We assess the practical validity of our approach, first by showing that our results apply in the framework of persistence-based shape classification via the ATOL procedure described in [34]. At last, numerical experiments are carried out on simulated and real datasets, encompassing text classification and large-scale graph classification.

## 1. Introduction

This paper handles the case where we observe $n$ i.i.d measures $X_1, \ldots, X_n$ on $\mathbb{R}^d$, 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 [33], repartition of clusters of diseases [16], modelisation of crime repartition [36] 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 [21], as well as topological data analysis, where persistence diagrams naturally appear as discrete measures in $\mathbb{R}^2$ [8, 10].

The objective of the paper is the following: we want to build from data an embedding of the sample measures into a finite-dimensional Euclidean space that preserves clusters structure, if any. Such an embedding, combined with standard clustering techniques should result in efficient measure clustering procedures. It is worth noting that 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 [33], genetics [34, 2], graphs clustering [7, 20] and shapes clustering [8].

Our vectorization scheme is quite simple: for a $k$-points vector $\mathbf{c} = (c_1, \ldots, c_k)$ in $(\mathbb{R}^d)^k$, we map each measure $X_i$ on $\mathbb{R}^d$ into a vector $v_i \in \mathbb{R}^k$ that roughly encodes how much weight $X_i$ spreads around every $c_j$. Then, in a mixture distribution framework, we provide general conditions on the $k$-points vectors that allow an efficient clustering based on the corresponding vectorization. Thus, efficiency of our measure clustering procedure essentially depends on our ability to find discriminative $k$-points vectors from sample.

Choosing a fixed grid of $\mathbb{R}^d$ as vectorization points is possible, however the dimension of such a vectorization is constrained and can grow quite large. Another choice could be to draw vectorization points at random from the sample measures. Note however that such a method could miss discriminative areas. We choose an intermediate approach, by building a $k$-points approximation of a central measure of $X_1, \ldots, X_n$. Note that finding compact representations of central tendencies of measures is an interesting problem in itself, especially when the sample measures are naturally organized around a central measure of interest, for instance in image analysis [14] or point processes modeling [33, 16, 36].

In [14], 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 $\nu$ 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 $X_1, \ldots, X_n$ in practice is too costly for large $n$'s, even with approximating algorithms [14, 31]. 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 $\mathbb{E}[X(A)]$ 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 $\mathbf{c} = (c_1, \ldots, c_k)$ that approximates well $\mathbb{E}(X)$, based on $X_1, \ldots, X_n$. To this aim, we introduce two algorithms (batch and mini-batch) that extend classical quantization techniques intended to solve the $k$-means problem [27]. In fact, these algorithms are built to solve the $k$-means problem for $\bar{X}_n$. We prove in Section 3.1 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 prove optimality of the classical quantization techniques [27, 25] in the point sample case.

We assess the overall validity of our appproach by giving theoretical guarantees on our vectorization and clustering process in a topological data analysis framework, where measures are persistence diagrams from different shapes. In

this case, vectorization via evaluation onto a fixed grid is the technique exposed in [2], whereas our method has clear connections with the procedures described in [44, 34]. Our theoretical results are given for vectorizations via evaluations of kernel functions around each point $c_j$, for a general class of kernel functions that encompasses the one used in [34]. Up to our knowledge, this provides the only theoretical guarantee on a measure-based clustering algorithm.

At last, we perform numerical experiments in Section 4 to assess the effectiveness of our method on real and synthetic data, in a clustering and classification framework. In a nutshell, our vectorization scheme combined with standard algorithms provides state-of-the art performances on various classification and clustering problems, with a lighter computational cost. The classification problems encompass sentiment analysis of IMDB reviews [26] as well as large-scale graph classification [35, 41]. Surprisingly, our somehow coarse approach turns out to outperform more involved methods in several large-scale graph classification problems.

The paper is organized as follows: Section 2 introduces our general vectorization technique, and conditions that guarantee a correct clustering based on it. Section 3 gives some detail on how our vectorization points are build, by introducing the mean measure $k$-points quantization problem. Then, two theoretically grounded algorithms are described to solve this problem from the sample $X_1, \ldots, X_n$. Section 3.3 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. The numerical experiments are exposed in Section 4. Sections 5 and 6 gather the main proofs of the results. Proofs of intermediate and technical results are deferred to Appendix A.

## 2. Vectorization and clustering of measures

Throughout the paper we will consider finite measures on the $d$-dimensional ball $\mathcal{B}(0, R)$ of the Euclidean space $\mathbb{R}^d$, equipped with its Borel algebra. Let $\mathcal{M}(R, M)$ denote the set of such measures of total mass smaller than $M$. For $\mu \in \mathcal{M}(R, M)$ and $f$ a borelian function from $\mathbb{R}^d$ to $\mathbb{R}$, we denote by $\int f(y)\mu(dy)$ integration of $f$ with respect to $\mu$, whenever $\int |f(y)|\mu(dy)$ is finite. We let $X$ be a random variable with values in $\mathcal{M}(R, M)$ (equipped with the smallest algebra such that $\int f dX$ is measurable, for any continuous and bounded $f$), and $X_1, \ldots, X_n$ be an i.i.d sample drawn from $X$.

### 2.1. Vectorization of measures

We aim to provide a map $v$ from $\mathcal{M}(R, M)$ into a finite-dimensional space that preserves separation between sources. That is, if we assume that there exists $(Z_1, \ldots, Z_n) \in [\![1, L]\!]^n$ a vector of (hidden) label variables such that $X_i \mid \{Z_i = \ell\} \sim X^{(\ell)}$, for different mixture components $X^{(1)}, \ldots, X^{(L)}$ taking values in $\mathcal{M}(R, M)$, we want $\|v(X_{i_1}) - v(X_{i_2})\|$ to be large whenever $Z_{i_1} \neq Z_{i_2}$.

The intuition is the following. Let $\mu_1 \neq \mu_2$ be two different measures on $\mathcal{B}(0, R)$. Then there exists $x \in \mathcal{B}(0, R)$ and $r > 0$ such that $\mu_1(\mathcal{B}(x, r)) \neq \mu_2(\mathcal{B}(x, r))$. Thus, the vectorization that sends $X_i$ into $X_i(\mathcal{B}(x, r))$ will allow to separate the sources $X^{(1)} = \delta_{\mu_1}$ and $X^{(2)} = \delta_{\mu_2}$. We extend this intuition to the multi-classes case. To adopt the quantization terminology, for $k \in \mathbb{N}^*$ we let $c_1, \ldots, c_k$ be *codepoints*, and $\mathbf{c} = (c_1, \ldots, c_k) \in (\mathbb{R}^d)^k$ is called a *codebook*. For a given codebook $\mathbf{c}$ and a scale $r > 0$, we may represent a measure $\mu$ via the vector of weights $(\mu(\mathcal{B}(c_1, r)), \ldots, \mu(\mathcal{B}(c_k, r)))$ that encodes the mass that $\mu$ spreads around every pole $c_j$. Provided that the codepoints are discriminative (this will be discussed in the following section), separation between clusters of measure will be preserved. In practice, convolution with kernels is often preferred to local masses (see, e.g., [34]). To ease computation, we will restrict ourselves to the following class of kernel functions.

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

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

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

Note that a $(p, \delta)$-kernel is also a $(q, \delta)$-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 [34] 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, a 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 $\psi$ is fixed, and, for a $k$-points codebook $\mathbf{c}$ and scale factor $\sigma$, consider the vectorization

$$v_{\mathbf{c},\sigma} : \begin{cases} \mathcal{M}(R, M) & \to & [0, M]^k \\ \mu & \mapsto & \left( \int \psi(\|u - c_1\|/\sigma)\mu(du), \ldots, \int \psi(\|u - c_k\|/\sigma)\mu(du) \right). \end{cases}$$
(1)

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 clustering purpose is the aim of the following section.

### 2.2. Discriminative codebooks and clustering of measures

In this section we assume that there exists a vector $(Z_1, \ldots, Z_n) \in [\![1, L]\!]^n$ of (hidden) label variables, and let $M_1, \ldots, M_L$ be such that, if $Z_i = \ell$, $X_i \in \mathcal{M}(R, M_\ell)$. We also denote $M = \max_{\ell \leq L} M_\ell$, and investigate under which conditions the vectorization exposed in the above section provides a representation

that is suitable for clustering. For a given codebook $\mathbf{c}$, we introduce the following definition of $(p, r, \Delta)$-scattering to quantify how well $\mathbf{c}$ will allow to separate clusters.

**Definition 2.** *Let $(p, r, \Delta) \in (\mathbb{N}^* \times \mathbb{R}^+ \times \mathbb{R}^+)$. A codebook $\mathbf{c} \in \mathcal{B}(0, R)^k$ is said to $(p, r, \Delta)$ -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*

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

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 3.** *Assume that $\mathbf{c} \in \mathcal{B}(0, R)^k$ shatters $X_1, \ldots, X_n$, with parameters $(p, r, \Delta)$. Then, if $\Psi$ is a $(p, \delta)$-kernel, with $\delta \leq \frac{\Delta}{4M}$, we have, for all $i_1$, $i_2 \in [\![1, n]\!]$ and $\sigma \in [r, 2r]$,*

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

A proof of Proposition 3 can be found in Section 5.1. This proposition sheds some light on how $X_1, \ldots, X_n$ has to be shattered with respect to the parameters of $\Psi$. 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 $\delta$ small enough compared to $1/M$, and thus $p$ large enough if $\Psi$ 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 $r$ then drives the choice of the bandwidth $\sigma$. Section 3.3 below provides an instance of shattered measures in the case where measures are persistence diagrams originating from well separated shapes. If the requirements of Proposition 3 are fulfilled, then a standard hierarchical clustering procedure such as Single Linkage with $L_\infty$ distance will separate the clusters for the scales smaller than $\Delta/2$. Note that, without more assumptions, this global clustering scheme can result in more than $L$ clusters, nested into the ground truth label classes.

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. For $p \in [\![1, +\infty]\!]$ and $\mu_1$, $\mu_2 \in \mathcal{M}(R, M)$ such that $\mu_1(\mathbb{R}^d) = \mu_2(\mathbb{R}^d)$, we denote by $W_p(\mu_1, \mu_2)$ the $p$-Wasserstein distance between $\mu_1$ and $\mu_2$.

**Definition 4.** *The sample of measures $X_1, \ldots X_n$ is called w-concentrated if, for all $i_1$, $i_2$ in $[\![1, n]\!]$ such that $Z_{i_1} = Z_{i_2}$,*

$$i) \quad X_{i_1}(\mathbb{R}^d) = X_{i_2}(\mathbb{R}^d), \qquad ii) \quad W_1(X_{i_1}, X_{i_2}) \leq w.$$

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 5.

**Proposition 5.** *Assume that $X_1, \ldots, X_n$ is w-concentrated. If $\Psi$ 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 $X_1, \ldots, X_n$ is $(p, r, \Delta)$-shattered by $\mathbf{c}$, and $(r\Delta/4)$-concentrated, then, for any $(p, \delta)$-kernel satisfying $\delta \leq \frac{\Delta}{4M}$, we have, for $\sigma \in [r, 2r]$,*

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

A proof of Proposition 5 is given in Section 5.2. An immediate consequence of Proposition 5 is that $(p, r, \Delta)$-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 $\sigma$ is crucial. Some heuristic is proposed in [34] in the special case of i.i.d persistence diagrams. An alternative calibration strategy is proposed in Section 3.3. Constructing shattering codebooks from sample is the purpose of the following section.

## 3. Quantization of the mean measure

To find discriminative codepoints from a measure sample $X_1, \ldots, X_n$, a naive approach could be to choose a $\varepsilon$-grid of $\mathcal{B}(0, R)$, for $\varepsilon$ small enough. Such an approach ignores the sample, and results in quite large codebooks, with $k \sim (R/\varepsilon)^d$ codepoints. An opposite approach is to sample $k_i$ points from every $X_i$, with $k = \sum_{i=1}^n k_i$, resulting in a $k$-points codebook that might miss some discriminative areas, especially in the case where discriminative areas have small $X_i$-masses. We propose an intermediate procedure, based on quantization of the mean measure. Definition 6 below introduces the mean measure.

**Definition 6.** *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*

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

*The empirical mean measure $\bar{X}_n$ may be defined via $\bar{X}_n(A) = \frac{1}{n} \sum_{i=1}^n X_i(A)$.*

In the case where the measures of interest are persistence diagrams, the mean measure defined above is the expected persistence diagram, defined in [11]. 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 $\mathcal{M}(R, M)$.

For a codebook $\mathbf{c} = (c_1, \ldots, c_k) \in \mathcal{B}(0, R)^k$, we let

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

so that $(W_1(\mathbf{c}), \ldots, W_k(\mathbf{c}))$ forms a partition of $\mathbb{R}^d$ and $N(\mathbf{c})$ represents the skeleton of the Voronoi diagram associated with $\mathbf{c}$. To grid the support of $\mathbb{E}(X)$, we want to minimize the distortion $F$, that is

$$F(\mathbf{c}) = \int \min_{j=1,\ldots,k} \|u - c_j\|^2 \mathbb{E}(X)(du) = W_2^2(\mathbb{E}(X), P_{\mathbf{c}}),$$

where $P_{\mathbf{c}} = \sum_{j=1}^k \mathbb{E}(X)(W_j(\mathbf{c}))\delta_{c_j}$ is a measure supported by $\mathbf{c}$. It is worth noting that if $P'_k = \sum_{j=1}^k \mu_j \delta_{c_j}$, with $\sum_{j=1}^k \mu_j = \mathbb{E}(X)(\mathbb{R}^d)$, then $W_2^2(\mathbb{E}(X), P'_k) \ge W_2^2(\mathbb{E}(X), P_{\mathbf{c}})$, so that $P_{\mathbf{c}}$ is the best approximation of $\mathbb{E}(X)$ with support $\mathbf{c}$ in terms of $W_2$ distance.

According to [18, Corollary 3.1], since $\mathbb{E}(X) \in \mathcal{M}(R, M)$, there exist minimizers $\mathbf{c}^*$ of $F(\mathbf{c})$, and we let $\mathcal{C}_{opt}$ denote the set of such minimizers. As well, if $\mathcal{M}_k(R, M)$ denotes the subset of $\mathcal{M}(R, M)$ that consists of distributions supported by $k$ points and $\mathbf{c}^* \in \mathcal{C}_{opt}$, then $P_{\mathbf{c}^*}$ is a best approximation of $\mathbb{E}(X)$ in $\mathcal{M}_k(R, M)$ in terms of $W_2$ distance. Note that distortions and thus optimal codebooks may be defined for $W_p$, $p \ne 2$. However, simple algorithms are available to approximate such a $\mathbf{c}^*$ in the special case $p = 2$, as discussed below.

### 3.1. *Batch and mini-batch algorithms*

This section exposes two algorithms that are intended to approximate a best $k$-points codebook for $\mathbb{E}(X)$ from a measure sample $X_1, \ldots, X_n$. These algorithms are extensions of two well-known clustering algorithms, namely the Lloyd algorithm ([25]) and Mac Queen algorithm ([27]). We first introduce the counterpart of Lloyd algorithm.

ALGORITHM 1. *Batch algorithm (Lloyd)*

```
Input : X_1, …, X_n and k ;
# Initialization
 Sample c_1^(0), c_2^(0),…c_k^(0) from X̄_n.;
while  c^(t+1) ≠ c^(t)  do:
  # Centroid update.
  for j in 1..k:
```

$$c_j^{(t+1)} = \frac{1}{\bar{X}_n(W_j(\mathbf{c}^{(t)}))} \int u \mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u) \bar{X}_n(du) \, ;$$

**Output** : $\mathbf{c}^{(T)}$ (codebook of the last iteration) .

Note that Algorithm 1 is a batch algorithm, in the sense that every iteration needs to process the whole data set $X_1, \ldots, X_n$. Fortunately, Theorem 9 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 [27] for standard $k$-means clustering is the following mini-batch algorithm. We let $\pi_{\mathcal{B}(0,R)^k}$ denote the projection onto $\mathcal{B}(0, R)^k$.

ALGORITHM 2. *Mini-batch algorithm (Mac-Queen)*

**Input** : $X_1, \ldots, X_n$, divided into mini-batches $(B_1, \ldots, B_T)$ with sizes $(n_1, \ldots, n_T)$, and $k$.
For $j = 1, \ldots, T$, $B_j$ is divided in two halves, $B_j^{(1)}$ and $B_j^{(2)}$.
Maximal radius R ;
```
# 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** $1 .. k$:

$$c_j^{(t+1)} = \pi_{\mathcal{B}(0,R)^k} \left( c_j^{(t)} - \frac{\int (c_j^{(t)} - u) \mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u) \bar{X}_{B_{t+1}^{(2)}}(du)}{(t+1)\bar{X}_{B_{t+1}^{(1)}}(W_j(\mathbf{c}^{(t)}))} \right) \, ;$$

**Output** : $\mathbf{c}^{(T)}$ (codebook of the last iteration) .

Note that the mini-batches split in two halves is motivated by technical considerations only that are exposed in Section 6.2. Whenever $n_i = 1$ for $i = 1, \ldots, n$ (and $B_i^{(1)} = B_i^{(2)} = \{X_i\}$), Algorithm 2 is a slight modification of the original Mac-Queen algorithm [27]. 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 $\hat{p}_j^{(t)} = \bar{X}_{B_t^{(1)}}(W_j(\mathbf{c}^{(t)}))$. These modifications are motivated by Theorem 10, that guarantees near-optimality of the output of Algorithm 2, provided that the mini-batches are large enough.

### 3.2. *Theoretical guarantees*

We now investigate the distortions of the outputs of Algorithm 1 and 2. In what follows, $F^*$ will denote the optimal distortion achievable with $k$ points, that is $F^* = F(\mathbf{c})$, where $\mathbf{c} \in \mathcal{C}_{opt}$. Note that $F^* = 0$ if and only if $\mathbb{E}(X)$ is supported by less than $k$ points, in which case the quantization problem is trivial. In what follows we assume that $\mathbb{E}(X)$ is supported by more than $k$ points. Basic properties of $\mathcal{C}_{opt}$ are recalled below.

**Proposition 7.** *[24, 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. $p_{min} = \inf_{\mathbf{c}^* \in \mathcal{C}_{opt}, j=1,\ldots,k} \mathbb{E}(X) \left(W_j(\mathbf{c}^*)\right) > 0$.

We will further assume that $\mathbb{E}(X)$ satisfies a so-called *margin condition*, defined in [22, Definition 2.1] and recalled below. For any subset $A$ of an Euclidean space, and $r > 0$, we denote by $\mathcal{B}(A, r)$ the set $\cup_{u \in A} \mathcal{B}(u, r)$.

**Definition 8.** $\mathbb{E}(X) \in \mathcal{M}(R, M)$ *satisfies a margin condition with radius $r_0 > 0$ if and only if, for all $0 \leq t \leq r_0$,*

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

In a nutshell, a margin condition ensures that the mean distribution $\mathbb{E}(X)$ is well-concentrated around $k$ poles. For instance, finitely-supported distributions satisfy a margin condition. Following [24], a margin condition will ensure that usual $k$-means type algorithms are almost optimal in terms of distortion. Up to our knowledge, margin-like conditions are always required to guarantee convergence of Lloyd-type algorithms [37, 24].

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 [24, Proposition 7]. Namely, for any empirically designed codebook $\hat{\mathbf{c}}$, we have

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

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 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 9, that investigates the performance of the output of Algorithm 1. We recall that, for $N \in \mathbb{N}^*$, $\mathcal{M}_N(R, M)$ denotes the subset of measures in $\mathcal{M}(R, M)$ that are supported by $N$ points.

**Theorem 9.** *Let $X \in \mathcal{M}_{N_{max}}(R, M)$, for some $N_{max} \in \mathbb{N}^*$. 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 $1 - 9e^{-c_1 n p_{min}^2 \kappa_0^2 / M^2} - e^{-x}$, where $c_1$ is a constant, we have*

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

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

A proof of Theorem 9 is given in Section 6.1. Combined with (2), Theorem 9 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. Note that Theorem 9 is valid for discrete measures that are supported by a uniformly bounded number of points $N_{max}$. This assumption is relaxed for Algorithm 2 in Theorem 10.

In the vector quantization case, Theorem 9 might be compared with [22, Theorem 3.1] for instance. In this case, the dependency on the dimension $d$ provided by Theorem 9 is sub-optimal. Slightly anticipating, dimension-free bounds in the mean-measure quantization case exist, for instance by considering the output of Algorithm 2.

Theorem 9 guarantees that choosing $T = 2\log(n)$ and repeating several Lloyd algorithms starting from different initializations provides an optimal quantization scheme. From a theoretical viewpoint, for a fixed $R_0$, the number of trials to get a good initialization grows exponentially with the dimension $d$. In fact, Algorithm 1 seems quite sensitive to the initialization step, as for every EM-like algorithm. From a practical viewpoint, we use 10 initializations using a $k$-`means++` sampling [3] for the experiments of Section 4, leading to results that are close to the reported ones for Algorithm 2.

Note that combining [24, Theorem 3] or [37] and a deviation inequality for distortions such as in [22] gives an alternative proof of the optimality of Lloyd type schemes, in the sample points case where $X_i = \delta_{X_i^{(1)}}$. In addition, Theorem 9 provides 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 6.1, 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 10.** *Let $X \in \mathcal{M}(R, M)$. 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 $(B_1, \ldots, B_T)$ 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*

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

*In the sample point case, the same result holds with the centroid update*

$$c_j^{(t+1)} = c_j^{(t)} - \frac{\int (c_j^{(t)} - u)\mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u)\bar{X}_{B_{t+1}}(du)}{(t+1)\bar{X}_{B_{t+1}}\left(W_j(\mathbf{c}^{(t)})\right)},$$

*that is without splitting the batches.*

A proof of Theorem 10 is given in Section 6.2. Note that Theorem 10 does not require the values of $X$ to be finitely supported measures, contrary to Theorem 9. Theorem 10 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, several initializations may be tried and the codebook with the best empirical distortion is chosen. In practice, the results of Section 4 are given for a single initialization at random using $k$-means++ sampling [3].

Note that Theorem 10 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., [6, Theorem 6.2]). In the point sample case, more refined bounds can be obtained, using for instance [22, 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 10 does not depend on the dimension $d$. This is also the case in [22, 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 9 and 10, 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.

We may now connect Theorem 9 and 10 to the results of Section 2.2.

**Corollary 11.** *Assume that $\mathbb{E}(X)$ satisfies the requirements of Theorem 9, and that $\mathbf{c}^*$ provides a $(p, r, \Delta)$ shattering of $X_1, \ldots, X_n$, with $p \geq 2$. Let $\hat{\mathbf{c}}_n$ denote the output of Algorithm 1. Then $\hat{\mathbf{c}}_n$ is a $(\lfloor \frac{p}{2} \rfloor, r, \Delta)$ shattering of $X_1, \ldots, X_n$, with probability larger than $1 - \exp\left[-C\left(\frac{nr^2 p_{min}^2}{p^2 M^2 R^2 k^2 d \log(k)} - \frac{p_{min}^2 B^2 r_0^2}{M^2 R^4 k^2 d \log(k)}\right)\right]$, where $C$ is a constant.*

A proof of Corollary 11 is given in Section 6.3. Corollary 11 ensures that if $\mathbf{c}^*$ shatters well $X_1, \ldots, X_n$, so will $\hat{\mathbf{c}}_n$, where $\hat{\mathbf{c}}_n$ is built with Algorithm 1. 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 describes such a situation.

### 3.3. Application: clustering persistence diagrams

In this section we investigate the properties of our measure-clustering scheme in a particular instance of i.i.d. measure observations. Recall that in our mixture model, $(Z_1, \ldots, Z_n)$ is the vector of i.i.d. labels such that $X_i \mid \{Z_i = \ell\} \sim X^{(\ell)}$. We further denote by $\pi_\ell = \mathbb{P}(Z_i = \ell)$, so that $X \sim \sum_{\ell=1}^L \pi_\ell X^{(\ell)}$.

Our mixture of persistence diagrams model is the following. For $\ell \in [\![1, L]\!]$ let $S^{(\ell)}$ denote a compact $d_\ell$-manifold and $D^{(\ell)}$ denote the persistence diagram generated by $d_{S^{(\ell)}}$, the distance to $S^{(\ell)}$. This persistence diagram can either be thought of as a multiset of points (set of points with integer-valued multiplicity) in the half-plane $H^+ = \{(b, d) \in \mathbb{R}^2 \mid 0 \leq b \leq d\}$ (for a general introduction to persistence diagrams the reader is referred to [5, Section 11.5]), or as a discrete measure on $H^+$ (see, e.g., [11] for more details on the measure point of view on

persistence diagrams). For a fixed scale $s > 0$, we define $D_{\geq s}^{(\ell)}$ as the thresholded persistence diagram of $d_{S^{(\ell)}}$, that is

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

where the multiplicities $n(b,d) \in \mathbb{N}^*$, and the $m_j^{(\ell)}$'s satisfy $(m_j^{(\ell)})_2 - (m_j^{(\ell)})_1 \geq s$. The following lemma ensures that $k_0^{(\ell)}$ is finite.

**Lemma 12.** *reach $\tau$. Let $S$ be a compact subset of $\mathbb{R}^d$, and $D$ denote the persistence diagram of the distance function $d_S$. 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 Appendix A.2. Now, for $\ell \in [\![1, L]\!]$ we let $\mathbb{Y}_{N_\ell}$ be an $N_\ell$-sample drawn on $S^{(\ell)}$ with density $f^{(\ell)}$ satisfying $f^{(\ell)}(u) \geq f_{min,\ell}$, for $u \in S^{(\ell)}$. If $\hat{D}^{(\ell)}$ denotes the persistence diagram generated by $d_{\mathbb{Y}_{N_\ell}}$, we define $X^{(\ell)}$ as a thresholded version of $\hat{D}^{(\ell)}$, that is

$$X^{(\ell)} \sim \hat{D}_{\geq s-h_\ell}^{(\ell)},$$

for some bandwidth $h_\ell$. To sum up, conditionally on $Z_i = \ell$, $X_i$ is a thresholded persistence diagram corresponding to a $N_\ell$-sampling of the shape $S^{(\ell)}$.

To retrieve the labels $Z_i$'s from a vectorization of the $X_i$'s, we have to assume that the persistence diagrams of the underlying shapes differ by at least one point.

**Definition 13.** *The shapes $S^{(1)}, \ldots, S^{(\ell)}$ are **discriminable at scale 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 $m_{\ell_1,\ell_2}$ satisfies the discrimination condition stated above, then $m_{\ell_1,\ell_2} \in D_{\geq s}^{(\ell_1)}$ or $m_{\ell_1,\ell_2} \in D_{\geq s}^{(\ell_2)}$. Next, we must ensure that optimal codebooks of the mean measure have codepoints close enough to discrimination points.

**Proposition 14.** *Let $h_\ell = \left(\frac{C_{d_\ell} \log(N_\ell)}{f_{min,\ell} N_\ell}\right)^{1/d_\ell}$, for some constants $C_{d_\ell}$, 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 $S^{(1)}, \ldots, S^{(L)}$ are discriminable at scale $s$, and let $m_1, \ldots, m_{k_0}$ denote the discrimination points. Let $K_0(h)$ denote*

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

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

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

The proof of Proposition 14 is given in Section 6.4. 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 14 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 $S^{(1)}, \ldots, S^{(L)}$ are discriminable at scale $s$ and that $k$ is large enough, we can prove that an optimal $k$-points codebook $\mathbf{c}^*$ is a $(p, r, \Delta)$-shattering of the sample, with high probability.

**Proposition 15.** *Assume that the requirements of Proposition 14 are satisfied. Let $\tilde{B} = \min_{i=1,\ldots,k_0, j=1,\ldots,K_0, j \neq i} \|m_i - m_j\|_\infty \wedge s$. Let $\kappa > 0$ be a small enough constant. Then, if $N_\ell$ is large enough for all $\ell \in [\![1, \ell]\!]$, $X_1, \ldots, X_n$ 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 15 is given in Section 6.5. In turn, Proposition 15 can be combined with Proposition 5 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 [34], $\psi_{AT} = x \mapsto \exp(-x)$.

**Corollary 16.** *Assume that the requirements of Proposition 15 are satisfied, and that $\mathbb{E}(X)$ satisfies a margin condition. For short, let $v_i$ denote the vectorization of $X_i$ based on the output of Algorithm 1. Then, with probability larger than*

$$1 - \exp\left[-C\left(\frac{nr^2 p_{min}^2}{p^2 M^2 R^2 k^2 d \log(k)} - \frac{p_{min}^2 B^2 r_0^2}{M^2 R^4 k^2 d \log(k)}\right)\right]$$

$$- n \max_{\ell \leq L} N_\ell^{-\left(\frac{(\kappa \tilde{B})^{d_\ell} f_{min,\ell} N_\ell}{C_\ell d_\ell \log(N_\ell)}\right)},$$

*where $\kappa$ and $C$ are small enough constants, we have*

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

*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 16 is given in Section 6.6. Corollary 16 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 16, Single Linkage with norm $\|.\|_\infty$ will provide an exact clustering. The probability bound in Corollary 16 shed some light on the quality of sampling of each shape that is required to achieve a perfect classification: roughly, for $N_\ell$ 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.

Investigating whether the margin condition required by Corollary 16 is satisfied in the mixture of shapes framework is beyond the scope of the paper. This condition is needed to prove that the algorithms exposed in Section 3.1 provide provably good approximations of optimal $\mathbf{c}^*$'s (those in turn can be proved to be discriminative codebooks). Experimental results below asses the validity of our algorithms in practice.

## 4. Experimental results

This section investigates numerical performances of our vectorization and clustering procedure. The vectorization step is carried out using the ATOL procedure (Automatic Topologically-Oriented Learning, [34]): codebooks are defined using Algorithm 2, and an exponential kernel $\Psi_{AT}$ is used to construct the embedding as in (1). As a remark, Algorithm 1 combined with other kernels such as $\psi_0$ lead to similar results in our three experiments.

It is worth noting that our method requires few to no tuning: for all experiments we use minibatches of size 1000 and the same calibration scheme (use a random 10% of the sample measures for learning the space quantization). We use the same kernel $\Psi_{AT}$, only the ATOL budget $k$ (length of the vectorization) will sometimes vary in the three experiments.

Our numerical experiments encompass synthetic measure clustering, but also investigate the performance of our vectorization technique on two (supervised) classification problems: large-scale graph classification and text classification. On these problems, we achieve state-of-the-art performances.

### *4.1. Measure clustering*

In this first experiment, we settle in the framework of Section 2.2. Namely, we consider an i.i.d sample drawn from a mixture of measures, and evaluate our method on a clustering task. We generate a synthetic mixture of $L$ measures. We first choose $p - 1$ centers shared by all mixture components on the $d - 1$-dimensional sphere of radius 10 and denote them by $\{S_d^i \mid i = 1, \ldots, p-1\}$; then for each mixture component another center is placed at a separate vertex of the $d$-dimensional unit cube (implying $L \leqslant 2^d$), labeled $\mathrm{Cube}_{\ell|d}$ for $\ell \in [\![1, \ldots, L]\!]$, so that the inner center is the only center that varies amongst mixture components, see instances Figure 1. The *support centers* $C^\ell$ for the mixture component $\ell$ are given by

$$C^\ell = \{S_d^i \mid i = 1, \ldots, p - 1\} \cup \mathrm{Cube}_{\ell|d}. \tag{3}$$

For every mixture component and signal level $r > 0$, we then make $N$ normal simultaneous draws with variance 1 centered around every element of $r \times C^\ell$, resulting in a point cloud of cardinality $p \times N$ that is interpreted as a measure. To sum up, the $\ell$-th mixture component $X^k$ has distribution

$$X^\ell = \bigcup_{c \in C^\ell} \bigcup_{i=1,\ldots,N} \{rc + \varepsilon_{c,i}\},$$

where the $\varepsilon$'s are i.i.d standard $d$-dimensional Gaussian random variables.

We compare with several measure clustering procedures, some of them being based on alternative vectorizations. Namely, the `rand` procedure randomly chooses codepoints among the sample measures points, whereas the `grid` procedure takes as codepoints a regular grid of $[0, 10r]^d$. For fair comparison we use the same kernel $\Psi_{AT}$ as in ATOL to vectorize from these two codebooks.

We also compare with two other standard clustering methods: the `histogram` procedure, that learns a regular tiling of the space and vectorizes measures by counting inside the tiles, and the `W2-spectral` method that is just spectral clustering based on the $W_2$ distance (to build the Gram matrix). Note that the `histogram` procedure is conceptually close to the `grid` procedure.

At last, for all methods the final clustering step consists in a standard k-means clustering with 100 initializations.

We designed these mixtures to exemplify problems where the data are significantly separated (by $L$ discriminative points corresponding to hypercube vertices), but also very similar by including draws from the same set of centers on the outer sphere. In this experiment $r$ may be thought of as the strength of the signal separation, and $p$ can be interpreted as a measure of noise. Empirically we use two different setups for the number of supporting centers $p$: $p = 4$ or $p = 20$, that we interpret respectively as low and high noise situations in Figure 1. We fix $L = 3$, and generate measures in dimension 2 and 5. We also set $N = 25$ and sample every mixture component 20 times, so that the final sample of measures consist of 60 i.i.d. finite measures, each with $25 \times p$ points.
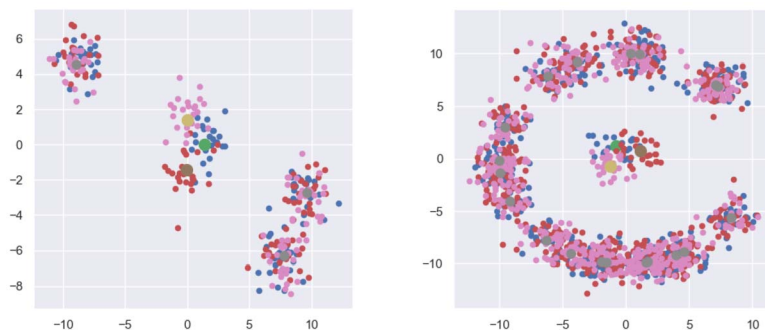
FIG 1. *Synthetic 3-mixture of measures instance (blue, dark brown and pink) and their support centers for indication (dark green for the outer centers, yellow, green and brown for the middle centers) with either 4 (left) or 20 (right) support centers in dimension 2, $r = 1$. The sets of points similarly coloured constitute measures.*

Regardless of the dimension, for the $p = 4$ setup, `W2-spectral` takes about 4.3 seconds on average, 0.2 seconds for ATOL, and the other methods are naturally faster. Note that for the $p = 20$ case, the `W2-spectral` clustering procedure runs takes too long to compute and is not reported, whereas ATOL runs in 0.5 seconds on average. Each experiment is performed identically a hundred times and the resulting average *normalized mutual information* (NMI) with true labels and 95% confidence intervals are presented for each method. Two sets of experiments are performed, corresponding to the following two questions.

**Q1**: at a given signal level $r = 1$, for increasing budget (that is, the size of the codebooks used in the vectorization process), how accurate are methods at clustering 3-mixtures? Q1 aims at comparing the various methods potentials at capturing the underlying signal with respect to budget.

Results are displayed in Figure 2. The ATOL procedure is shown to produce the best results in almost all situations. Namely, it is the procedure that will most likely cluster the mixture exactly while requiring the least amount of budget for doing so. The grid procedures `grid` and `histogram` naturally suffer from the dimension growth (bottom panels). The oscillations of the `histogram` procedure performances, exposed in the upper right panel, may be due to the existence of a center tile (that cannot discriminate between several vertices of the cube) that our oversimple implementation allows. To draw a fair comparison, the local minima of the `histogram` curve are not considered as representative, leading to an overall flat behavior of this method in the case $d = 2, p = 20$. In dimension 2 the `grid` procedure shows to be very efficient, but for measures in dimension 5 it seems always better to select codepoints at random, rather than using some instance of regular grid. Whenever possible, `W2-spectral` outperforms all vectorization-based methods, at the price of a significantly larger computational cost. Note however that the performance gap with ATOL is discernable for small budgets only ($k \leq 15$).

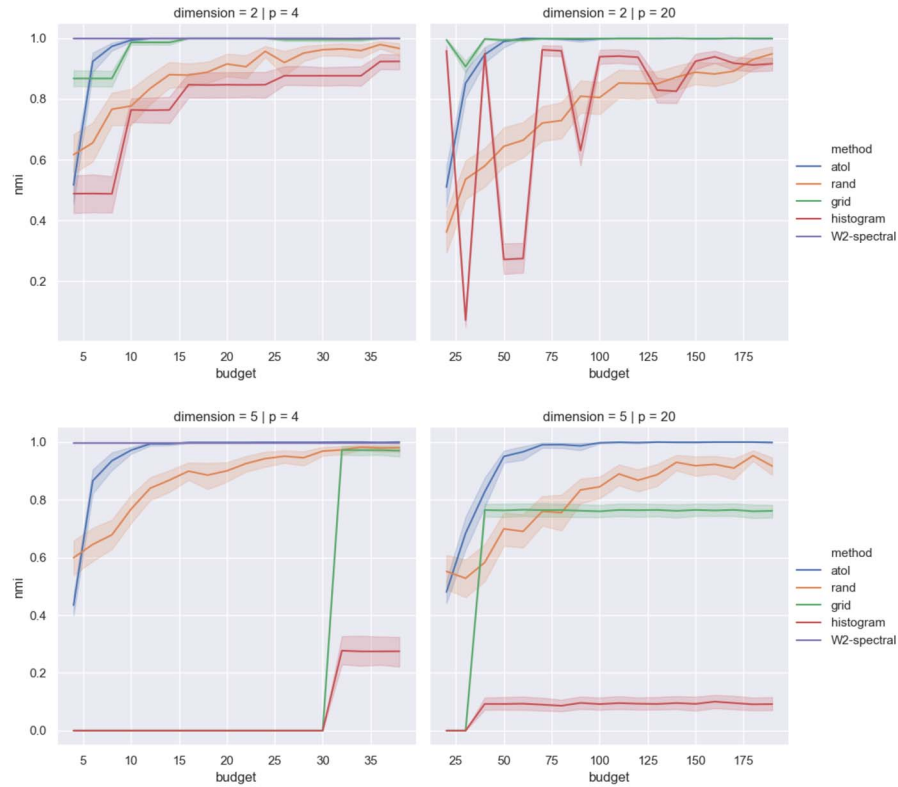**Q2**: for a given budget $k = 32$ and increasing signal level, how accurate are

FIG 2. *NMI as budget increases, $r = 1$, increasing dimensions along rows, low noise and high noise contexts (left and right columns)*

methods at clustering 3-mixtures? Q2 investigates how the various methods fare with respect to signal strength at a given, low codebook length $k = 32$, the most favorably low budget for regular grid procedures in dimension 5.

Results are shown in Figure 3. ATOL and `grid` procedures give similar results, and compare well to `W2-spectral` in the case $p = 4$. Figure 3 also confirms the intuition that the `rand` procedure is more sensitive to $p$ than competitors, resulting in significantly weaker performances when $p = 20$.

### 4.2. Large-scale graph classification

We now look at the large-scale graph classification problem. Provided that graphs may be considered as measures, our vectorization method provides an embedding into $\mathbb{R}^k$ that may be combined with standard learning algorithms. In this framework, our vectorization procedure can be thought of as a dimensionality reduction technique for supervised learning, that is performed in an unsupervised way. Note that the notion of shattering codebook introduced in
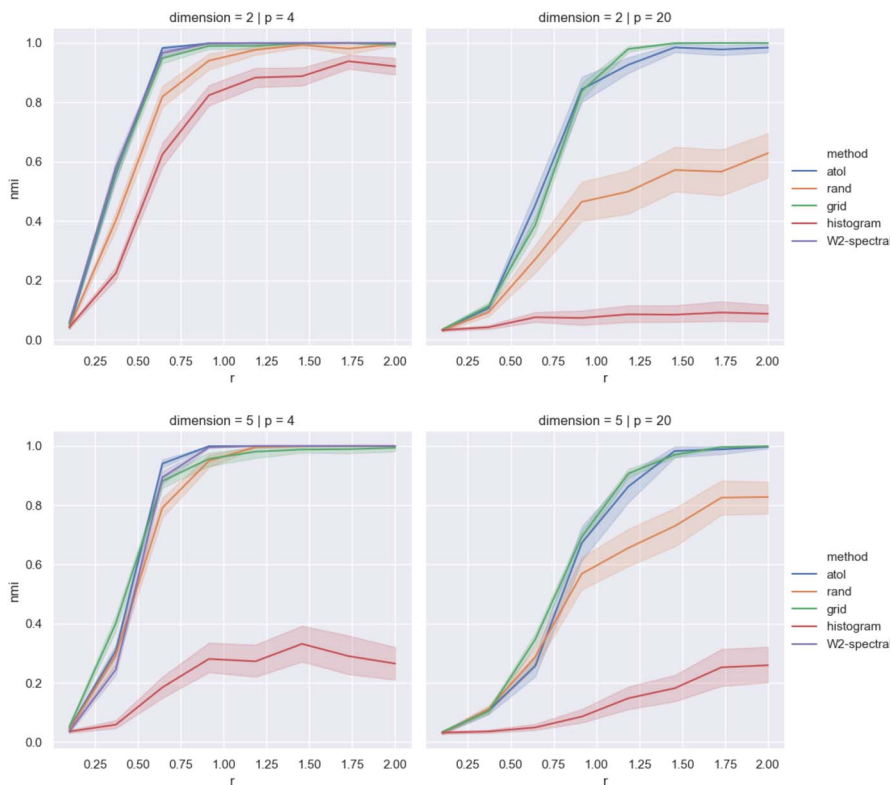
FIG 3. *NMI as r increases, budget k = 32, increasing dimensions along rows, low noise and high noise contexts (left and right columns)*

Section 2.2 is still relevant in this context, since a $(p, r, \Delta)$-shattering codebook would lead to exact sample classification if combined with a classification tree for instance. Theoretically investigating the generalization error of such an approach is beyond the scope of the paper.

As exposed above, since ATOL is an unsupervised procedure that is not specifically conceived for graphs, it is likely that other dedicated techniques would provide better performances in this graph classification problem. However, its simplicity can be a competitive advantage for large-scale applications.

There are multiple ways to interpret graphs as measures. In this section we borrow from [7]: for a diffusion time $t > 0$ we compute the Heat Kernel Signatures $(\mathrm{HKS}_t)$ for all vertices in a graph, so that each graph $G(V, E)$ is embedded in $\mathbb{R}^{|V|}$ (see, e.g., [7, Section 2.2]). Then four graph descriptors per diffusion time may be computed using the extended persistence framework (see, e.g., [7, Section 2.1]), that is four types of persistence diagrams (PDs). Schematically for a

TABLE 1

*Large graph binary classification problems. Mean ROC-AUC and standard deviations.*

| method | | SF | NetLSD | FGSD | GeoScat | ATOL |
|---|---|---|---|---|---|---|
| problem | size | [15] | [38] | [40] | [19] | |
| `reddit threads` | (203K) | $81.4\pm.2$ | $\mathbf{82.7\pm.1}$ | $82.5\pm.2$ | $80.0\pm.1$ | $80.7\pm.1$ |
| `twitch egos` | (127K) | $67.8\pm.3$ | $63.1\pm.2$ | $\mathbf{70.5\pm.3}$ | $69.7\pm.1$ | $69.7\pm.1$ |
| `github stargazers` | (12.7K) | $55.8\pm.1$ | $63.2\pm.1$ | $65.6\pm.1$ | $54.6\pm.3$ | $\mathbf{72.3\pm.4}$ |
| `deezer ego nets` | (9.6K) | $50.1\pm.1$ | $52.2\pm.1$ | $\mathbf{52.6\pm.1}$ | $52.2\pm.3$ | $51.0\pm.6$ |

graph $G(V, E)$ the descriptors are derived as:

$$G(V, E) \xrightarrow[\text{signatures}]{\text{heat kernel}} \text{HKS}_t(G) \in \mathbb{R}^{|V|},$$

$$G(V, E) \xrightarrow[\text{persistence}]{\text{extended}} \text{PD}(\text{HKS}_t(G), G) \in (\mathcal{M}(\mathbb{R}^2))^4. \tag{4}$$

In these experiments we will show that the graph embedding strategy of (4) paired with ATOL can perform up to the state of the art on large-scale graph classification problems. We will also demonstrate that ATOL is well assisted but not dependent on the aforementioned TDA tools.

**Large-scale binary classification from [35]**

Recently [35] introduced large-scale graph datasets of social or web origin. For each dataset the associated task is binary classification. The authors perform a 80% train/test split of the data and report mean area under the curve (AUC) along with standard errors over a hundred experiments for all the following graph embedding methods. SF from [15] is a simple graph embedding method that extracts the $k$ lowest spectral values from the graph Laplacian, and a standard Random Forest Classifier (RFC) for classification. NetLSD from [38] uses a more refined representation of the graph Laplacian, the heat trace signature (a global variant of the HKS) of a graph using 250 diffusion times, and a 1-layer neural network (NN) for classification. FGSD from [40] computes the biharmonic spectral distance of a graph and uses histogram vectorization with small binwidths that results in vectorization of size $[100, 1000000]$, with a Support Vector Machine (SVM) classifier. GeoScattering from [19] uses graph wavelet transform to produce 125 graph embedding features, also with a SVM classifier.

We add our own results for ATOL paired with (4): we use the extended persistence diagrams as input for ATOL computed from the HKS values with diffusion times $t_1 = .1, t_2 = 10$, and vectorize the diagrams with budget $k = 10$ for each diagram type and diffusion time so that the resulting vectorization for graph $G$ is $v_{\text{ATOL}}(G) \in \mathbb{R}^{2 \times 4 \times 10}$. We then train a standard RFC (as in [15], we use the implementation from `sklearn` [30] with all default parameters) on the resulting vectorized measures.

The results are shown Table 1. ATOL is close to or over the state-of-the-art for all four datasets. Most of these methods operate directly from graph Laplacians so they are fairly comparable, in essence or as for the dimension of the embedding that is used. The most positive results on `github stargazers` improves on the best method by more than 6 points.

**A variant of the large-scale graph classification from [34]**

The graph classification tasks above were binary classifications. [41] introduced now popular datasets of large-scale graphs associated with multiple classes. These datasets have been tackled with top performant graph methods including the graph kernel methods RetGK from [42], WKPI from [43] and GNTK from [17] (combined with a graph neural network), and the aforementioned graph embedding method FGSD from [40]. PersLay from [7] is not designed for graphs but used (4) as input for a 1-NN classifier. Lastly, in [34], the authors also used (4) and the same diagrams as input data for the ATOL procedure and obtained competitive results within reasonable computation times (the largest is 110 seconds, for the COLLAB dataset, see [34, Section 3.1] for all computation times).

Here we propose to bypass the topological feature extraction step in (4). Instead of topological descriptors, we use simpler graph descriptors: we compute four HKS descriptors corresponding to diffusion times $t_1 = .1, t_2 = 1, t_3 = 10, t_4 = 100$ for all vertices in a graph, but this time directly interpret the output as a measure embedding in dimension 4. From there we use ATOL with $k = 80$ budget. Therefore each graph $G(V, E)$ is embedded in $\mathbb{R}^{4|V|}$ seen as $\mathcal{M}(\mathbb{R}^4)$ and our measure vectorization framework is readily applicable from there. To sum-up we now use the point of view:

$$G(V, E) \xrightarrow[\text{signatures}]{\text{heat kernel}} \text{HKS}_{t_1, t_2, t_3, t_4}(G) \in \mathbb{R}^{4|V|} \approx \mathcal{M}(\mathbb{R}^4). \tag{5}$$

We do not claim that this particular embedding is the correct way to handle graphs in general, much to the contrary: instead we simply show that our methodology can handle all sort of embedding strategies, with good results to support it for the exposed experiments. It is important to note that the proposed transformation from graphs to measures can be computed using any type of node or edge embedding. Our vectorization method could be applied the same way as for HKS's.

TABLE 2
*Mean accuracies and standard deviations for the large multi-classes problems of [41].*

| method | RetGK | FGSD | WKPI | GNTK | PersLay | ATOL | ATOL |
|---|---|---|---|---|---|---|---|
| problem | [42] | [40] | [43] | [17] | [7] | with (4) [34] | with (5) |
| REDDIT (5K, 5 classes) | $56.1\pm.5$ | $47.8$ | $59.5\pm.6$ | — | $55.6\pm.3$ | $\mathbf{67.1\pm.3}$ | $\mathbf{66.1\pm.2}$ |
| REDDIT (12K, 11 classes) | $48.7\pm.2$ | — | $48.5\pm.5$ | — | $47.7\pm.2$ | $\mathbf{51.4\pm.2}$ | $\mathbf{50.7\pm.3}$ |
| COLLAB (5K, 3 classes) | $81.0\pm.3$ | $80.0$ | — | $83.6\pm.1$ | $76.4\pm.4$ | $\mathbf{88.3\pm.2}$ | $\mathbf{88.5\pm.1}$ |
| IMDB-B (1K, 2 classes) | $71.9\pm1.$ | $73.6$ | $75.1\pm1.1$ | $\mathbf{76.9\pm3.6}$ | $71.2\pm.7$ | $74.8\pm.3$ | $73.9\pm.5$ |
| IMDB-M (1.5K, 3 classes) | $47.7\pm.3$ | $\mathbf{52.4}$ | $48.4\pm.5$ | $\mathbf{52.8\pm4.6}$ | $48.8\pm.6$ | $47.8\pm.7$ | $47.0\pm.5$ |

On Table 2 we quote results and competitors from [34] and on the right column we add our own experiment with ATOL. The ATOL methodology works very efficiently with the direct HKS embedding as well, although the results tend to be slightly inferior. This may hint at the fact that although PDs are not essential to capturing signal from this dataset, they can be a significant addition for doing so. Overall ATOL with (5), though much lighter than ATOL with (4), remains competitive for large-scale graph multiclass classification.

### *4.3. Text classification with word embedding*

At last we intend to apply our methodology in a high-dimensional framework, namely text classification. Basically, texts are sequences of words and if one forgets about word order, a text can be thought of as a measure in the (fairly unstructured) space of words. We use the `word2vec` word embedding technique ([29]), that uses a two-layer neural network to learn a real-valued vector representation for each word in a dictionary in such a way that distances between words are learnt to reflect the semantic closeness between them, with respect to the corpus. Therefore, for a given dimension $E \in \mathbb{N}$, every word $w$ is mapped to the embedding space $v_{\texttt{word2vec}}(w) \in \mathbb{R}^E$, and we can use word embedding to interpret texts as measures in a given word embedding space:

$$T \in \mathcal{M}(\{\text{words}\}) \xrightarrow[\text{embedding}]{\text{word}} T_{\texttt{word2vec}} = \big[v_{\texttt{word2vec}}(w)\big]_{w \in T} \in \mathcal{M}(\mathbb{R}^E). \quad (6)$$

In practice we will use the Large Movie Review Dataset from [26] for our text corpus, and learn word embedding on this corpus. To ease the intuition, Figure 4 below depicts the codepoints provided by Algorithm 2, based on a 2-dimensional word embedding. Note that this word embedding step does not depend on the classification task that will follow.
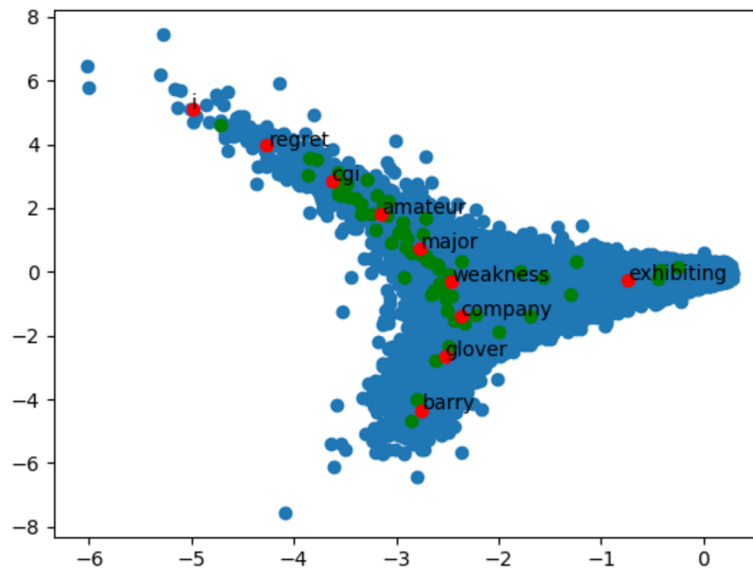


FIG 4. *Blue: point cloud embedding in dimension 2 for the words of all reviews; green: point cloud embedding for the words of the first review ("Story of a man who has unnatural feelings for a pig. [...]"); red: codepoints computed by Algorithm 2, with the text print of their closest corresponding word.*

We now turn to the task of classification and use the binary sentiment annotation (positive or negative reviews) from the dataset. For classifying texts with

modern machine-learning techniques there are several problems to overcome: important words in a text can be found at almost any place, texts usually have different sizes so that they have to be padded in some way, etc. The measure viewpoint is a direct and simple solution.

The successful way to proceed after a word embedding step is to use some form of deep neural network learning. We learn a 100-dimensional word embedding using the `word2vec` implementation from the `gensim` module, then compare the following classification methods: directly run a recurrent neural network (LSTM with 64 units), against the Atol vectorization followed by a simple dense layer with 32 units. We measure accuracies through a single 10-fold.

The Atol vectorization with word embedding in dimension 100, 20 centers, and 1-NN classifier reaches 85.6 accuracy with .95 standard deviation. The average quantization and vectorization times are respectively 5.5 and 208.3 seconds. The recurrent neural network alternative with word embedding in dimension 100 reaches 89.3 mean accuracy with .44 standard deviation, for an average run time of about 1 hour.

Naturally the results are overall greatly hindered by the fact that we have treated texts as a separate collection of words, forgetting sentence structure when most competitive methods use complex neural networks to analyse $n$-uplets of words as additional inputs. Additional precision can be also gained using a task-dependend embedding of words, at the price of a larger computation time (see for instance the `kaggle` winner algorithm [12], with precision 99% and run time 10379 seconds).

## 5. Proofs for Section 2

### *5.1. Proof of Proposition 3*

Assume that $X_1, \ldots, X_n$ is $(p, r, \Delta)$-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

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

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

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

On the other hand, we have that

$$
\begin{aligned}
\int \Psi(\|u - c_1\|/\sigma) X_{i_2}(du) &\leq \int \Psi(\|u - c_1\|/\sigma) \mathbb{1}_{\mathcal{B}(c_1, 4pr)} X_{i_2}(du) \\
&\quad + \int \Psi(\|u - c_1\|/\sigma) \mathbb{1}_{(\mathcal{B}(c_1, 4pr))^c} X_{i_2}(du)
\end{aligned}
$$

$$\leq X_{i_2}(\mathcal{B}(c_1, 4pr)) + \delta X_{i_2}((\mathcal{B}(c_1, 4pr))^c)$$
$$\leq X_{i_1}(\mathcal{B}(c_1, r/p)) - \Delta + \delta M.$$

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

### 5.2. Proof of Proposition 5

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

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

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. Proofs for Section 3

### 6.1. Proof of Theorem 9

Throughout this section we assume that $\mathbb{E}(X)$ satisfies a margin condition with radius $r_0$, and that $\mathbb{P}(X \in \mathcal{M}_{N_{max}}(R, M)) = 1$. The proof of Theorem 9 is based on the following lemma, which 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 17.** *Assume that* $\mathbf{c}^{(0)} \in \mathcal{B}(\mathbf{c}^*, R_0)$. *Then, with probability larger than* $1 - 9e^{-c_1 n p_{min}/M} - e^{-x}$, *for* $n$ *large enough, we have, for every* $t$,

$$\|\mathbf{c}^{(t+1)} - \mathbf{c}^*\|^2 \leq \frac{3}{4}\|\mathbf{c}^{(t)} - \mathbf{c}^*\|^2 + \frac{K}{p_{min}^2}D_n^2, \qquad (7)$$

*where* $D_n = \frac{CRM}{\sqrt{n}}\left( k\sqrt{d\log(k)} + \sqrt{x} \right)$ *and* $C, K$ *are positive constants.*

The proof of Lemma 17 is deferred to Appendix A.1.1.

*Proof of Theorem 9.* Equipped with Lemma 17, the proof of Theorem 9 easily follows. On the probability event of Lemma 17, using (7) we have that

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

### 6.2. Proof of Theorem 10

The proof of Theorem 10 follows the proof of [32, Lemma 1]. Throughout this section we assume that $\mathbb{E}(X)$ satisfies a margin condition with radius $r_0$. The proof of Theorem 10 is based on the following lemma. Recall that $R_0 = \frac{Br_0}{16\sqrt{2}R}$, $\kappa_0 = \frac{R_0}{R}$. We let $\mathbf{c}^{(t)}$ denote the output of Algorithm 2 (or its variant in the sample point case).

**Lemma 18.** *Assume that* $\mathbf{c}^{(0)} \in \mathcal{B}(\mathbf{c}^*, R_0)$, *and* $n_t \geq c_0 \frac{k^2 M^2}{p_{min}^2 \kappa_0^2} \log(n)$, *for some constant* $c_0$, *with* $n \geq k$. *Then we have, for any* $t = 0, \ldots, T - 1$,

$$\mathbb{E}\left(\|\mathbf{c}^{(t+1)} - \mathbf{c}^*\|^2\right) \leq \left(1 - \frac{2 - K_1}{t + 1}\right)\mathbb{E}\left(\|\mathbf{c}^{(t)} - \mathbf{c}^*\|^2\right) + \frac{16kMR^2}{p_{min}(t+1)^2}, \quad (8)$$

*with* $K_1 \leq 0.5$.

The proof of Lemma 18 is deferred to Appendix A.1.2.

*Proof of Theorem 10.* Equipped with Lemma 18, we can prove Theorem 10 using the same method as in the proof of [32, Lemma 1]. Namely, denoting by $a_t = \|\mathbf{c}^{(t)} - \mathbf{c}^*\|^2$, we prove recursively that

$$\mathbb{E}a_t \leq \frac{32kMR^2}{p_{min}t}.$$

Denote by $G = \frac{16kMR^2}{p_{min}}$. The case $t = 1$ is obvious. Next, assuming that $\mathbb{E}a_t \leq \frac{2G}{t}$ and using (8) we may write

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

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

### 6.3. Proof of Corollary 11

For $n$ large enough, with probability larger than

$$1 - \exp\left[-C\left(\frac{nr^2 p_{min}^2}{p^2 M^2 R^2 k^2 d \log(k)} - \frac{p_{min}^2 B^2 r_0^2}{M^2 R^4 k^2 d \log(k)}\right)\right],$$

we have $\|\hat{\mathbf{c}}_n - \mathbf{c}^*\| \leq \frac{r}{p}$, according to Theorem 9. 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 $X_{i_1}\left(\mathcal{B}(\hat{c}_1, 2r/p)\right) \geq X_{i_1}\left(\mathcal{B}(\hat{c}_1, r/p)\right))$. Noting that $X_{i_2}\left(\mathcal{B}(\hat{c}_1, 4(p/2)r)\right) \leq X_{i_2}\left(\mathcal{B}(c_1^*, 4pr)\right)$ entails that

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

### 6.4. Proof of Proposition 14

The proof of Proposition 14 relies on the two following lemmas. First, Lemma 19 below ensures that the persistence diagram transformation is stable with respect to the sup-norm.

**Lemma 19.** *[13] If $X$ and $Y$ are compact sets of $\mathbb{R}^D$, then*

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

To apply Lemma 19 to $d_{\mathbb{Y}_{N_\ell}}$ and $d_S^{(\ell)}$, for $\ell \in [\![1, L]\!]$, a bound on $\|d_{\mathbb{Y}_{N_\ell}} - d_S^{(\ell)}\|_\infty$ may be derived from the following result.

**Lemma 20.** *[1, Lemma B.7] Let $S \subset \mathbb{R}^D$ be a $d$-dimensional submanifold with positive reach $\tau_S$, 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 $S$. Assume that for all $x \in S$, $0 < f_{min} \leq f(x)$, 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 $h \leq \tau_S/4$, then, with probability larger than $1 - \left(\frac{1}{N}\right)^{\frac{k}{d}}$, we have*

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

We let $Z$ denote the latent label variable, so that $X \mid \{Z = \ell\} \sim X^{(\ell)}$, or equivalently $X = X^{(Z)}$. Now, for $\ell \in [\![1, L]\!]$, we let $\alpha_\ell = \left(\frac{1}{N_\ell}\right)^{d_\ell + \frac{2}{d_\ell}}$, $h_\ell = \left(\frac{C_{d_\ell} \log(N_\ell)}{f_{min,\ell} N_\ell}\right)^{1/d_\ell}$, for some constants $C_{d_\ell}$, and $A = \{\|d_{Y_{N_Z}} - d_{S_Z}\|_\infty > h_Z\}$, 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 $\bigcup_{\ell=1}^L D_{\geq s}^{(\ell)} \setminus \{m_1, \ldots, m_{k_0}\} \subset \bigcup_{s=1}^{K_0(h)} \mathcal{B}_\infty(m_{k_0+s}, h)$, and $\mathbf{m} = (m_1, \ldots, m_{k_0 + K_0(h)})$. At last, we let $R = \max_{\ell \leq L} diam(S_\ell)$. Choosing the $N_\ell$'s large enough so that $D_{\geq s - h_\ell}^{(\ell)} = D_{\geq s}^{(\ell)}$ and $s/2 > h_\ell$, $X^{(Z)}$ and $D_{\geq s}^{(Z)}$ have the same number of (multiple) points, when $A^c$ occurs. Further, we have

$$F(\mathbf{m}) = \mathbb{E}\left(\sum_{\ell=1}^L \mathbb{1}_{Z=\ell} \int \min_{j=1,\ldots,k_0+K_0(h)} \|u - m_j\|^2 X^{(\ell)}(du)\right)$$

$$= \mathbb{E}\left(\sum_{\ell=1}^L \mathbb{1}_{Z=\ell \cap A} \int \min_{j=1,\ldots,k_0+K_0(h)} \|u - m_j\|^2 X^{(\ell)}(du)\right)$$

$$+ \mathbb{E}\left(\sum_{\ell=1}^L \mathbb{1}_{Z=\ell \cap A^c} \int \min_{j=1,\ldots,k_0+K_0(h)} \|u - m_j\|^2 X^{(\ell)}(du)\right),$$

so that

$$F(\mathbf{m}) \leq \mathbb{E}\left(\sum_{\ell=1}^L \mathbb{1}_{Z=\ell \cap A} 4R^2 N^{d_\ell}\right) + \mathbb{E}\left(\sum_{\ell=1}^L \mathbb{1}_{Z=\ell \cap A^c} M^{(\ell)} 2h_\ell^2\right)$$

$$\leq 2h^2\bar{M} + 4R^2 \sum_{\ell=1}^{L} \pi_\ell \alpha_\ell N^{d_\ell}.$$

For $N_\ell$ large enough so that $\alpha_\ell N_\ell^{d_\ell} \leq \frac{\bar{M}h_\ell^2}{R^2}$, we have $F(\mathbf{m}) \leq 6h^2\bar{M}$.

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

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

$$\geq \mathbb{E}\left( \sum_{\ell=1}^{L} \mathbb{1}_{A^c \cap Z=\ell} n^{(\ell)}(m_1) \left( 5\sqrt{\frac{\bar{M}}{\pi_{min}}} - 1 \right)^2 \right) h^2.$$

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

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

hence the result.

### 6.5. Proof of Proposition 15

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 14, we choose $N_\ell$, $\ell \leq L$ large enough so that $5\sqrt{\bar{M}}h/(\sqrt{\pi_{min}}) \leq \kappa\tilde{B}$. 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

$$\mathbb{P}\left( \exists i \in [\![1, n]\!] \mid d_B(X_i, D_{\geq s}^{Z_i}) > \kappa\tilde{B} \right) \leq \sum_{i=1}^{n} \mathbb{P}\left( d_B(X_i, D^{Z_i}) > \kappa\tilde{B} \right)$$

$$\leq \sum_{i=1}^{n} \sum_{\ell=1}^{L} \pi_\ell \alpha_\ell \leq n \max_{\ell \leq L} \alpha_\ell.$$

For the remaining of the proof we assume that, for $i = 1, \ldots, n$, $d_B(X_i, D_{\geq s}^{Z_i}) \leq \kappa\tilde{B}$, that occurs with probability larger than $1 - n \max_{\ell \leq L} \alpha_\ell$. Recall that on this probability event, under the assumptions of Proposition 14, $X_i$ and $D_{\geq s}^{Z_i}$ have the same number of (multiple) points. 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 $D_{\geq s}^{Z_{i_1}}(\{m_1\}) \geq D_{\geq s}^{Z_{i_2}}(\{m_1\}) + 1$. 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 $d_B(X_{i_1}, D_{\geq s}^{Z_{i_1}}) \leq \kappa\tilde{B} < h_0$, we get

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

On the other hand, since $4rp \leq (\frac{1}{2} - \kappa)\tilde{B}$, we also have $X_{i_2}(\mathcal{B}(c_1^*, 4rp)) = D_{\geq s}^{Z_{i_2}}(\{m_1\})$. Thus $X_1, \ldots, X_n$ is $(p, r, 1)$-shattered by $\mathbf{c}^*$.

### 6.6. Proof of Corollary 16

In the case where $\Psi = \Psi_{AT}$, we have that $\Psi_{AT}$ is a $(p, 1/p)$ kernel. The requirement $1/p \leq \frac{1}{4M}$ of Proposition 3 is thus satisfied for $p_{AT} = \lceil 4M \rceil$. On the other hand, choosing $r_{AT} = \frac{\tilde{B}}{32p_{AT}}$ ensures that $8r_{AT}p_{AT} \leq (1/2 - \kappa)\tilde{B}$ and $\frac{r_{AT}}{2p_{AT}} \geq 2\kappa\tilde{B}$, for $\kappa$ small enough. Thus, the requirements of Proposition 15 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 $\kappa$ small enough to conclude that $X_1, \ldots, X_n$ is $\frac{r_{AT}}{4}$-concentrated on the probability event described in Proposition 15. Thus Proposition 5 applies.

The case $\Psi = \Psi_0$ is simpler. Since $\Psi_0$ is a $(1, 0)$-kernel, we obviously have that $0 \leq \frac{1}{2M}$, so that the requirement of Proposition 15 is satisfied. With $p_0 = 1$ and $r_0 = \frac{\tilde{B}}{16}$ we immediately get that $r_0/(2p_0) \geq 2\kappa\tilde{B}$ and $8r_0p_0 \leq (1/2 - \kappa\tilde{B})$, for $\kappa$ 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 $\kappa$ small enough. Thus Proposition 5 applies.

### 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] ARTHUR, D. and VASSILVITSKII, S. (2007). k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 1027–1035. ACM, New York. MR2485254

[4] BARTLETT, P. and LUGOSI, G. (1999). An inequality for uniform deviations of sample averages from their means. *Statist. Probab. Lett.* **44** 55–62. MR1706315

[5] BOISSONNAT, J.-D., CHAZAL, F. and YVINEC, M. (2018). *Geometric and Topological Inference* **57**. Cambridge University Press. MR3837127

[6] 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

[7] 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* 1904.09378.

[8] 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.

[9] CHAZAL, F., COHEN-STEINER, D. and LIEUTIER, A. (2009). A sampling theory for compact sets in Euclidean space. *Discrete & Computational Geometry* **41** 461–479. MR2486371

[10] CHAZAL, F., DE SILVA, V., GLISSE, M. and OUDOT, S. (2016). *The structure and stability of persistence modules*. Springer International Publishing. MR3524869

[11] 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

[12] CHERNIUK, A. (2019). IMDB review challenge using Word2Vec and BiLSTM. https://www.kaggle.com/alexcherniuk/imdb-review-word2vec-bilstm-99-acc/notebook.

[13] COHEN-STEINER, D., EDELSBRUNNER, H. and HARER, J. (2007). Stability of persistence diagrams. *Discrete Comput. Geom.* **37** 103–120. MR2279866

[14] CUTURI, M. and DOUCET, A. (2013). Fast Computation of Wasserstein Barycenters. *arXiv e-prints* 1310.4375. MR3178606

[15] DE LARA, N. and PINEAU, E. (2018). A Simple Baseline Algorithm for Graph Classification.

[16] 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. MR3405479

[17] DU, S. S., HOU, K., SALAKHUTDINOV, R. R., POCZOS, B., WANG, R. and XU, K. (2019). Graph Neural Tangent Kernel: Fusing Graph Neural Networks with Graph Kernels. In *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox and R. Garnett, eds.) 5724–5734. Curran Associates, Inc.

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

[19] GAO, F., WOLF, G. and HIRN, M. (2019). Geometric Scattering for Graph Data Analysis. In *Proceedings of the 36th International Conference on Machine Learning* (K. CHAUDHURI and R. SALAKHUTDINOV, eds.). *Proceedings of Machine Learning Research* **97** 2122–2131. PMLR, Long Beach, California, USA.

[20] HOAN TRAN, Q., VO, V. T. and HASEGAWA, Y. (2018). Scale-variant topological information for characterizing the structure of complex networks. *arXiv e-prints* 1811.03573. MR4019475

[21] 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.

[22] LEVRARD, C. (2015). Nonasymptotic bounds for vector quantization in Hilbert spaces. *Ann. Statist.* **43** 592–619. MR3316191

[23] LEVRARD, C. (2015). Supplement to "Non-asymptotic bounds for vector quantization".

[24] LEVRARD, C. (2018). Quantization/Clustering: when and why does k-means work. *Journal de la Société Française de Statistiques* **159**. MR3803122

[25] LLOYD, S. P. (1982). Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28** 129–137. MR651807

[26] MAAS, A. L., DALY, R. E., PHAM, P. T., HUANG, D., NG, A. Y. and POTTS, C. (2011). Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* 142–150. Association for Computational Linguistics, Portland, Oregon, USA.

[27] 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

[28] MENDELSON, S. and VERSHYNIN, R. (2003). Entropy and the combinatorial dimension. *Invent. Math.* **152** 37–55. MR1965359

[29] MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S. and DEAN, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) 3111–3119. Curran Associates, Inc.

[30] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M. and DUCHESNAY, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12** 2825–2830. MR2854348

[31] 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.

[32] RAKHLIN, A., SHAMIR, O. and SRIDHARAN, K. (2011). Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. *arXiv e-prints* 1109.5647.

[33] 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.

[34] ROYER, M., CHAZAL, F., LEVRARD, C., IKE, Y. and UMEDA, Y. (2020). ATOL: Measure Vectorisation for Automatic Topologically-Oriented Learning. working paper or preprint.

[35] ROZEMBERCZKI, B., KISS, O. and SARKAR, R. (2020). An API Oriented Open-source Python Framework for Unsupervised Learning on Graphs.

[36] 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* 1701.05863. MR4085877

[37] 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.

[38] TSITSULIN, A., MOTTIN, D., KARRAS, P., BRONSTEIN, A. and MÜLLER, E. (2018). NetLSD. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[39] 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

[40] VERMA, S. and ZHANG, Z.-L. (2017). Hunt For The Unique, Stable, Sparse And Fast Feature Learning On Graphs. In *Advances in Neural Information Processing Systems* 88–98.

[41] YANARDAG, P. and VISHWANATHAN, S. V. N. (2015). Deep Graph Kernels. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '15* 1365–1374. Association for Computing Machinery, New York, NY, USA.

[42] ZHANG, Z., WANG, M., XIANG, Y., HUANG, Y. and NEHORAI, A. (2018). RetGK: Graph Kernels based on Return Probabilities of Random Walks. In *Advances in Neural Information Processing Systems* 3968–3978.

[43] ZHAO, Q. and WANG, Y. (2019). Learning metrics for persistence-based summaries and applications for graph classification. In *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox and R. Garnett, eds.) 9855–9866. Curran Associates, Inc.

[44] ZIELIŃSKI, B., LIPIŃSKI, M., JUDA, M., ZEPPELZAUER, M. and DŁOTKO, P. (2018). Persistence Bag-of-Words for Topological Data Analysis. *arXiv e-prints* 1812.09245.

## Appendix A: Technical proofs

To ease understanding the statements of the results are recalled before their proofs.

### A.1. Proofs for Section 6

A key ingredient of the proofs of Lemma 17 and 18 is the following Lemma 21, ensuring that around optimal codebooks the expected gradients of Algorithms 1 and 2 are almost Lipschitz.

**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*

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

The proof of Lemma 21 follows from [23, Section A.3].

#### A.1.1. Proof of Lemma 17

**Lemma** (17). *Assume that* $\mathbf{c}^{(0)} \in \mathcal{B}(\mathbf{c}^*, R_0)$. *Then, with probability larger than* $1 - 9e^{-c_1 n p_{min}/M} - e^{-x}$, *for* $n$ *large enough, we have, for every* $t$,

$$\|\mathbf{c}^{(t+1)} - \mathbf{c}^*\|^2 \leq \frac{3}{4}\|\mathbf{c}^{(t)} - \mathbf{c}^*\|^2 + \frac{K}{p_{min}^2}D_n^2, \qquad (7)$$

*where* $D_n = \frac{CRM}{\sqrt{n}}\left(k\sqrt{d\log(k)} + \sqrt{x}\right)$ *and* $C, K$ *are positive constants.*

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, for $j \in [\![1, k]\!]$,

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

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 Lemma 17 will make use of the following concentration lemma.

**Lemma 22.** *With probability larger than* $1 - 8e^{-x}$, *for all* $\mathbf{c} \in \mathcal{B}(0, R)^k$,

$$\hat{p}_j(\mathbf{c}) \leq p_j(\mathbf{c}) + \sqrt{\frac{4Mc_0 k d \log(k)\log(2nN_{max})}{n}} + \frac{4Mx}{n}\sqrt{p_j(\mathbf{c})}$$

$$\hat{p}_j(\mathbf{c}) \geq p_j(\mathbf{c}) - \frac{4Mc_0 k d \log(k)\log(2nN_{max})}{n} - \frac{4Mx}{n}$$
$$- \sqrt{\frac{4Mc_0 k d \log(k)\log(2nN_{max})}{n}} + \frac{4Mx}{n}\sqrt{p_j(\mathbf{c})},$$

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

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

*where $C$ is a constant.*

The proof of Lemma 22 is given in Appendix A.1.3, and is based on empirical processes theory.

*Proof of Lemma 17.* Let $\mathbf{c} \in \mathcal{B}(\mathbf{c}^*, R_0)$, and

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

We decompose $\|\hat{m}(\mathbf{c}) - \mathbf{c}^*\|^2$ as follows.

$$\|\hat{m}(\mathbf{c}) - \mathbf{c}^*\|^2 = \|\mathbf{c} - \mathbf{c}^*\|^2 + 2\langle \hat{m}(\mathbf{c}) - \mathbf{c}, \mathbf{c} - \mathbf{c}^* \rangle + \|\hat{m}(\mathbf{c}) - \mathbf{c}\|^2. \quad (9)$$

Next, we bound the first term of (9).

$$2\langle \hat{m}(\mathbf{c}) - \mathbf{c}, \mathbf{c} - \mathbf{c}^* \rangle = 2\sum_{j=1}^{k} \frac{1}{\hat{p}_j(\mathbf{c})} \left\langle \int (u - c_j) \mathbb{1}_{W_j(\mathbf{c})}(u)\bar{X}_n(du), c_j - c_j^* \right\rangle$$

$$\leq 2\sum_{j=1}^{k} \frac{1}{\hat{p}_j(\mathbf{c})} \left\langle \int (u - c_j) \mathbb{1}_{W_j(\mathbf{c})}(u)\mathbb{E}(X)(du), c_j - c_j^* \right\rangle$$

$$+ 2D_n \sqrt{\sum_{j=1}^{k} \frac{\|c_j - c_j^*\|^2}{\hat{p}_j(\mathbf{c})^2}}$$

$$\leq 2\sum_{j=1}^{k} \frac{1}{\hat{p}_j(\mathbf{c})} \left\langle p_j(\mathbf{c}^*)(c_j^* - c_j), c_j - c_j^* \right\rangle$$

$$+ \frac{2p_{min}}{8\sqrt{2}} \|\mathbf{c} - \mathbf{c}^*\| \sqrt{\sum_{j=1}^{k} \frac{\|c_j - c_j^*\|^2}{\hat{p}_j(\mathbf{c})^2}}$$

$$+ 2D_n \sqrt{\sum_{j=1}^{k} \frac{\|c_j - c_j^*\|^2}{\hat{p}_j(\mathbf{c})^2}},$$

where the last line follows from Lemma 21. Now, using Lemma 22 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

$$2\left\langle \hat{m}(\mathbf{c}) - \mathbf{c}, \mathbf{c} - \mathbf{c}^* \right\rangle \leq -2\sum_{j=1}^{k} \frac{p_j(\mathbf{c}^*)}{\hat{p}_j(\mathbf{c})}\|c_j - c_j^*\|^2 + \frac{32}{124\sqrt{2}}\|\mathbf{c} - \mathbf{c}^*\|^2$$

$$+ K_1\|\mathbf{c} - \mathbf{c}^*\|^2 + K_1^{-1}\frac{32^2}{31^2 p_{min}^2}D_n^2, \quad (10)$$

where $K_1 > 0$ is to be fixed later. Then, the second term of (9) may be bounded as follows.

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

$$= \sum_{j=1}^{k} \frac{\left\|p_j(\mathbf{c}^*)(c_j - c_j^*) + \Delta_j(\mathbf{c}) + \Delta_{n,j}(\mathbf{c})\right\|^2}{\hat{p}_j(\mathbf{c})^2},$$

where

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

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

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

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

$$\|\hat{m}(\mathbf{c}) - \mathbf{c}\|^2 \leq (1 + K_2 + K_3)\sum_{j=1}^{k} \frac{p_j(\mathbf{c}^*)^2}{\hat{p}_j(\mathbf{c})^2}\|c_j - c_j^*\|^2$$

$$+ \left(1 + K_2^{-1} + K_4\right)\sum_{j=1}^{k} \frac{\|\Delta_j(\mathbf{c})\|^2}{\hat{p}_j(\mathbf{c})^2}$$

$$+ \left(1 + K_3^{-1} + K_4^{-1}\right)\sum_{j=1}^{k} \frac{\|\Delta_{n,j}(\mathbf{c})\|^2}{\hat{p}_j(\mathbf{c})^2}$$

$$\leq (1 + K_2 + K_3)\sum_{j=1}^{k} \frac{p_j(\mathbf{c}^*)^2}{\hat{p}_j(\mathbf{c})^2}\|c_j - c_j^*\|^2$$

$$+ \left(1 + K_2^{-1} + K_4\right)\frac{32^2}{31^2 \times 128}\|\mathbf{c} - \mathbf{c}^*\|^2$$

$$+ \left(1 + K_3^{-1} + K_4^{-1}\right)\frac{32^2}{31^2 p_{min}^2}D_n^2, \quad (11)$$

where $K_2$, $K_3$ and $K_4$ are positive constants to be fixed later. Combining (10) and (11) yields that

$$\|\hat{m}(\mathbf{c}) - \mathbf{c}^*\|^2 \leq \|\mathbf{c} - \mathbf{c}^*\|^2 \left(1 + K_1 + \frac{32}{124\sqrt{2}} + \frac{32^2}{31^2 \times 128} \left(1 + K_2^{-1} + K_4\right)\right)$$

$$- 2 \sum_{j=1}^{k} \frac{p_j(\mathbf{c}^*)}{\hat{p}_j(\mathbf{c})} \|c_j - c_j^*\|^2$$

$$+ (1 + K_2 + K_3) \sum_{j=1}^{k} \frac{p_j(\mathbf{c}^*)^2}{\hat{p}_j(\mathbf{c})^2} \|c_j - c_j^*\|^2$$

$$+ D_n^2 \frac{32^2}{31^2 p_{min}^2} \left(1 + K_1^{-1} + K_3^{-1} + K_4^{-1}\right).$$

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

$$\|\hat{m}(\mathbf{c}) - \mathbf{c}^*\|^2 \leq \|\mathbf{c} - \mathbf{c}^*\|^2 \left(0.62 + K_1 + K_3 \frac{32^2}{31^2} + K_4 \frac{32^2}{31^2 \times 128}\right)$$

$$+ D_n^2 \frac{32^2}{31^2 p_{min}^2} \left(1 + K_1^{-1} + K_3^{-1} + K_4^{-1}\right)$$

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

for $K_1$, $K_3$ and $K_4$ small enough. Now, according to Lemma 22, 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.

Recalling that, for any $t$, $\mathbf{c}^{(t+1)} = \hat{m}(\mathbf{c}^{(t)})$, a straightforward recursion entails that, on a global probability event that has probability larger than $1 - 9e^{-cn\kappa_0^2 p_{min}^2 / M^2} - e^{-x}$, for $c$ small enough, provided that $\mathbf{c}^{(0)} \in \mathcal{B}(\mathbf{c}^*, R_0)$, we have for any $t \geq 0$ $\mathbf{c}^{(t)} \in \mathcal{B}(\mathbf{c}^*, R_0)$ and

$$\|\mathbf{c}^{(t+1)} - \mathbf{c}^*\|^2 \leq \frac{3}{4} \|\mathbf{c}^{(t)} - \mathbf{c}^*\|^2 + \frac{K}{p_{min}^2} D_n^2$$

$$\leq \frac{3}{4} \|\mathbf{c}^{(t)} - \mathbf{c}^*\|^2 + \frac{K}{p_{min}^2} \frac{C^2 R^2 M^2}{n} \left(k\sqrt{d\log(k)} + \sqrt{x}\right)^2. \quad \square$$

### A.1.2. *Proof of Lemma 18*

**Lemma** (18). *Assume that $\mathbf{c}^{(0)} \in \mathcal{B}(\mathbf{c}^*, R_0)$, and $n_t \geq c_0 \frac{k^2 M^2}{p_{min}^2 \kappa_0^2} \log(n)$, for some constant $c_0$, with $n \geq k$. Then we have, for any $t = 0, \ldots, T - 1$,*

$$\mathbb{E}\left(\|\mathbf{c}^{(t+1)} - \mathbf{c}^*\|^2\right) \leq \left(1 - \frac{2 - K_1}{t+1}\right)\mathbb{E}\left(\|\mathbf{c}^{(t)} - \mathbf{c}^*\|^2\right) + \frac{16kMR^2}{p_{min}(t+1)^2}, \qquad (8)$$

*with $K_1 \leq 0.5$.*

The proof of Lemma 18 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,*

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

A proof of Lemma 23 is given in Appendix A.1.4.

*Proof of Lemma 18.* Assume that $n \geq k$, and let $n_t = |B_t|/2 \geq \frac{CM}{p_{min}}\log(n)$, for $C$ large enough to be fixed later. For a given $t \leq T$, denote by $\hat{p}_j^{(t)} = \bar{X}_{B_t^{(1)}}(W_j(\mathbf{c}^{(t)}))$, and by $A_{t,1}$ and $A_{t,2}$ the events

$$A_{t,1} = \left\{\forall j = 1\ldots, k \quad |\hat{p}_j^{(t)} - p_j^{(t)}| \leq \frac{p_{min} + \sqrt{p_j^{(t)}p_{min}}}{256}\right\},$$

$$A_{t,2} = \left\{\forall j = 1\ldots, k \quad \left\|\int (c_j^{(t)} - u)\mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u)(\bar{X}_{B_t^{(2)}} - \mathbb{E}(X))(du)\right\|\right.$$

$$\left. \leq 8R\sqrt{\frac{Mkp_{min}}{C}}\right\}.$$

According to Lemma 23 with $x = 4\log(2n)$, we have $\mathbb{P}(A_{t,j}^c) \leq k/(2n^4)$, for $j = 1, 2$. We let also $A_{\leq t}$ denote the event $\cap_{r \leq t}A_{r_1} \cap A_{r_2}$, that has probability larger than $1 - kt/n^4$. First we prove that if $\mathbf{c}^{(0)} \in \mathcal{B}(\mathbf{c}^*, R_0)$, then, on $A_{\leq t}$, for all $r \leq t$, $\mathbf{c}^{(r)} \in \mathcal{B}(\mathbf{c}^*, R_0)$. We proceed recursively, assuming that $\mathbf{c}^{(t)} \in \mathcal{B}(\mathbf{c}^*, R_0)$. Then, on $A_{\leq t+1}$, 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 $g_{t+1} = \left(\frac{\int (c_j^{(t)} - u)\mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u)\bar{X}_{B_{t+1}^{(2)}}(du)}{\hat{p}_j^{t+1}}\right)_j$, the recursion equation entails that

$$a_{t+1} = \|\pi_{B(0,R)^k}\left(\mathbf{c}^{(t)} - \frac{g_{t+1}}{t+1}\right) - \mathbf{c}^*\|^2$$

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

$$(12)$$

As in the proof of Lemma 17, denote by

$$\Delta_j^t = \int (u - c_j^{(t)}) \mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u) \mathbb{E}(X)(du) - p_j^*(c_j^* - c_j^{(t)})$$

$$\Delta_{n,j}^{t+1} = \int (u - c_j^{(t)}) \mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u)(\bar{X}_{B_{t+1}^{(2)}} - \mathbb{E}(X))(du)$$

$$D_n^{t+1} = \sqrt{\sum_{j=1}^k \|\Delta_{n,j}^{t+1}\|^2}.$$

We have that

$$-\frac{2}{t+1}\langle g_{t+1}, \mathbf{c} - \mathbf{c}^* \rangle \le -\frac{2}{t+1}\sum_{j=1}^k \left( \frac{p_j^*}{\hat{p}_j^{t+1}}\|c_j^{(t)} - c_j^*\|^2 - \frac{\|\Delta_{j,n}^{t+1}\|\|c_j^{(t)} - c_j^*\|}{\hat{p}_j^{t+1}} \right.$$

$$\left. - \frac{\|c_j^{(t)} - c_j^*\|\|\Delta_j^t\|}{\hat{p}_j^{t+1}} \right)$$

$$\le -2\frac{32}{33(t+1)}\|\mathbf{c}^{(t)} - \mathbf{c}^*\|^2$$

$$+ \frac{64}{31p_{min}(t+1)}\|\mathbf{c}^{(t)} - \mathbf{c}^*\|D_n^{t+1}$$

$$+ \frac{64}{8\sqrt{2} \times 31(t+1)}\|\mathbf{c}^{(t)} - \mathbf{c}^*\|^2$$

$$\le \|\mathbf{c}^{(t)} - \mathbf{c}^*\|^2 \left( \frac{-64}{33(t+1)} + \frac{K_4}{t+1} + \frac{64}{8\sqrt{2} \times 31(t+1)} \right)$$

$$+ K_4^{-1} \left( \frac{32}{31p_{min}}D_n^{t+1} \right)^2,$$

according to Lemma 21, where $K_4$ denotes a constant. Next, the second term in (12) may be bounded by

$$\|g_{t+1}\|^2 \le \sum_{j=1}^k \frac{1}{(\hat{p}_j^{t+1})^2}(p_j^*)^2\|c_j^{(t)} - c_j^*\|^2(1 + K_1 + K_2)$$

$$+ \frac{p_{min}^2}{128\min_j (\hat{p}_j^{t+1})^2}\|\mathbf{c}^{(t)} - \mathbf{c}^*\|^2 \left(1 + K_2^{-1} + K_3\right)$$

$$+ \frac{1}{\min_j (\hat{p}_j^{t+1})^2}\left(1 + K_1^{-1} + K_3^{-1}\right)(D_n^{t+1})^2$$

$$\le \frac{32^2}{31^2}\|\mathbf{c}^{(t)} - \mathbf{c}^*\|^2 \left(1 + K_1 + K_2 + \frac{1 + K_2^{-1} + K_3}{128}\right)$$

$$+ (D_n^{t+1})^2 \frac{32^2(1 + K_1^{-1} + K_3^{-1})}{31^2 p_{min}^2},$$

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

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

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 $n_t \geq c_0 \frac{k^2 M^2}{p_{min}^2 \kappa_0^2} \log(n)$, $n \geq k$, where $c_0$ is an absolute constant, on $A_{\leq t+1}$ we have

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

Thus, provided that $\mathbf{c}^{(0)} \in \mathcal{B}(\mathbf{c}^*, R_0)$, on $A_{\leq t}$ we have $\mathbf{c}^{(p)} \in \mathcal{B}(\mathbf{c}^*, R_0)$, for $p \leq t$.

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

$$
\mathbb{E}a_{t+1} \leq \left(1 - \frac{2 - K_1}{t+1}\right)\mathbb{E}a_t + \frac{16kMR^2}{p_{min}(t+1)^2},
$$

where $K_1 < 0.5$. First, we may write

$$
\mathbb{E}a_{t+1} = \mathbb{E}\left(a_{t+1}\mathbb{1}_{A_{\leq t}}\mathbb{1}_{A_{t+1,1}}\right) + R_1,
$$

with $R_1 \leq (4k^2R^2)/n^3 \leq 4kMR^2/(p_{min}(t+1)^2)$, since $p_{min} \leq M/k$ and $t + 1 \leq T \leq n$. Then, using (12) entails

$$
\begin{aligned}
\mathbb{E}\left(a_{t+1}\mathbb{1}_{A_{\leq t} \cap A_{t+1,1}}\right) \leq\ & \mathbb{E}a_t - \frac{2}{t+1}\mathbb{E}\left(\left\langle g_{t+1}, \mathbf{c}^{(t)} - \mathbf{c}^*\right\rangle \mathbb{1}_{A_{\leq t} \cap A_{t+1,1}}\right) \\
& + \frac{1}{(t+1)^2}\mathbb{E}\left(\|g_{t+1}\|^2 \mathbb{1}_{A_{\leq t} \cap A_{t+1,1}}\right).
\end{aligned}
$$

Next, we bound the scalar product as follows.

$$\mathbb{E}\left(\left\langle -g_{t+1}, \mathbf{c}^{(t)} - \mathbf{c}^* \right\rangle \mathbb{1}_{A_{\leq t} \cap A_{t+1,1}}\right) =$$

$$\mathbb{E}\left(\mathbb{1}_{A_{\leq t}} \sum_{j=1}^{k} E_t\left(\frac{\left\langle \int (u - c_j^{(t)}) \mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u) \bar{X}_{B_{t+1}^{(2)}}(du), c_j^{(t)} - c_j^* \right\rangle}{\hat{p}_j^{t+1}} \mathbb{1}_{A_{t+1,1}}\right)\right),$$

where, for any $j \in [\![1, k]\!]$, we have

$$E_t\left(\frac{\left\langle \int (u - c_j^{(t)}) \mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u) \bar{X}_{B_{t+1}^{(2)}}(du), c_j^{(t)} - c_j^* \right\rangle}{\hat{p}_j^{t+1}} \mathbb{1}_{A_{t+1,1}}\right)$$

$$= E_t\left(\left\langle \int (u - c_j^{(t)}) \mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u) \bar{X}_{B_{t+1}^{(2)}}(du), c_j^{(t)} - c_j^* \right\rangle\right) E_t\left(\frac{1}{\hat{p}_j^{t+1}} \mathbb{1}_{A_{t+1,1}}\right).$$

The first term may be bounded by

$$E_t\left(\left\langle \int (u - c_j^{(t)}) \mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u) \bar{X}_{B_{t+1}^{(2)}}(du), c_j^{(t)} - c_j^* \right\rangle\right) \mathbb{1}_{A_{\leq t}}$$

$$\leq -p_j^* \|c_j^{(t)} - c_j^*\|^2 \mathbb{1}_{A_{\leq t}} + \|\Delta_{j,t}\| \|c_j^{(t)} - c_j^*\|,$$

and the second term by

$$\frac{32}{33 p_j^*} \mathbb{1}_{A_{\leq t}} \leq E_t\left(\frac{1}{\hat{p}_j^{t+1}} \mathbb{1}_{A_{t+1,1}}\right) \mathbb{1}_{A_{\leq t}} \leq \frac{32}{31 p_j^*}.$$

This gives

$$\mathbb{E}\left(\left\langle -g_{t+1}, \mathbf{c}^{(t)} - \mathbf{c}^* \right\rangle \mathbb{1}_{A_{\leq t} \cap A_{t+1,1}}\right) \leq -\frac{32}{33} \mathbb{E}\left(a_t \mathbb{1}_{A_{\leq t}}\right)$$

$$+ \frac{32}{31 p_{min}} \mathbb{E}\left(\sum_{j=1}^{k} \|\Delta_{j,t}\| \|c_j^{(t)} - c_j^*\| \mathbb{1}_{A_{\leq t}}\right)$$

$$\leq -\frac{32}{33} \mathbb{E}(a_t) + \frac{32}{31 \times 8\sqrt{2}} \mathbb{E}(a_t) + \frac{32}{33} \frac{4k^2 R^2}{n^3},$$

according to Lemma 21. At last, the bound on $\|g_{t+1}\|^2$ writes as follows.

$$\mathbb{E}\left(\|g_{t+1}\|^2 \mathbb{1}_{A_{\leq t} \cap A_{t+1,1}}\right) = \mathbb{E}\left(\mathbb{1}_{A_{\leq t}} E_t\left(\|g_{t+1}\|^2 \mathbb{1}_{A_{t+1,1}}\right)\right)$$

$$\leq \frac{32}{31} \frac{4kMR^2}{p_{min}}.$$

Gathering all pieces leads to

$$\mathbb{E} a_{t+1} \leq \left(1 - \frac{2 - K_1}{t+1}\right) \mathbb{E} a_t + \frac{16kMR^2}{p_{min}(t+1)^2},$$

with $K_1 \leq 0.5$.

At last, in the point sample case where we observe $n$ points in $\mathbb{R}^d$ $X_1, \ldots, X_n$ i.i.d with distribution $X$, recall that we take

$$\mathbf{c}^{(t+1)} = \mathbf{c}^{(t)} - \frac{g_{t+1}}{t+1}, \quad with \quad g_{t+1} = \frac{\int (c_j^{(t)} - u) \mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u) \bar{X}_{B_{t+1}}(du)}{\hat{p}_j^{t+1}}.$$

With a slight abuse of notation we denote by $A_{\leq t}$ the event onto which the concentration inequalities of Lemma 23 are satisfied (with $B_t^{(1)} = B_t^{(2)} = B_t$). The first step of the proof is the same: if $n_t \geq c_0 \frac{k}{p_{min}^2 \kappa_0^2}$ and $n \geq k$, then, on $A_{\leq t}$, for all $j \leq t$, $a_j \leq R_0$. It remains to prove the recursion inequality

$$\mathbb{E}a_{t+1} \leq \left(1 - \frac{2 - K_1}{t+1}\right) \mathbb{E}a_t + \frac{16kR^2}{p_{min}(t+1)^2}.$$

We proceed as before, writing $\mathbb{E}(a_{t+1}) = \mathbb{E}(a_{t+1}\mathbb{1}_{A_{\leq t}}) + R_1$, for $R_1$ that satisfies $R_1 \leq 4kR^2/(p_{min}(t+1)^2)$, and

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

Note that $\|g_{t+1}\|^2 \leq 4kR^2$, so that the second term might be bounded by $4kR^2/(t+1)^2$. To bound the scalar product, we proceed as follows. Let $j \in [\![1, k]\!]$, and, for $i \in B_{t+1}$, denote by $U_{i,j}$ the random variable $\mathbb{1}_{W_j(\mathbf{c}^{(t)})}(X_i)$. We have

$$\mathbb{E}\left(\frac{\left\langle \int (u - c_j^{(t)})\mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u)\bar{X}_{B_{t+1}}(du), c_j^{(t)} - c_j^* \right\rangle}{\hat{p}_j^{t+1}} \mathbb{1}_{A_{\leq t}}\right) =$$

$$\mathbb{E}\left(\frac{\mathbb{1}_{A_{\leq t}}}{\hat{p}_j^{t+1}}\left\langle \mathbb{E}\left[\int (u - c_j^{(t)})\mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u)\bar{X}_{B_{t+1}}(du) \mid \mathcal{F}_t, (U_{i,j})_{i \in B_{t+1}}\right],\right.$$

$$\left. c_j^{(t)} - c_j^* \right\rangle\right)$$

$$= \mathbb{E}\left(\frac{\mathbb{1}_{A_{\leq t}}}{n_{t+1}\hat{p}_j^{t+1}} \sum_{i \in B_{t+1}} U_{i,j}\left[-\frac{p_j^*}{p_j^t}\|c_j^{(t)} - c_j^*\|^2 + \frac{\left\langle \Delta_{j,t}, c_j^{(t)} - c_j^* \right\rangle}{p_j^t}\right]\right)$$

$$\leq -\frac{32}{33}\mathbb{E}(\|c_j^* - c_j\|^2\mathbb{1}_{A_{\leq t}}) + \frac{32}{31p_{min}}\mathbb{E}(\|\Delta_{j,t}\|\|c_j^t - c_j^*\|\mathbb{1}_{A_{\leq t}}).$$

The remaining of the proof is the same as for the sample measure case. □

### A.1.3. Proof of Lemma 22

Let $Z_1$ denote the process

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

Note that the VC dimension of Voronoi cells in a $k$-points Voronoi diagram is at most $c_0 k d \log(k)$ ([39, 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*

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

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

For the sake of completeness a proof of Lemma 24 is given in Appendix A.1.5. Next, introducing $\sigma_1, \ldots, \sigma_n$ independent Rademacher variables, we get

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{(P_n' - P_n)f}{\sqrt{(P_n'f + P_nf)/2}} \geq t\right) \leq \mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i(f(X_i) - f(X_i'))}{\sqrt{(P_n'f + P_nf)/2}} \geq t\right)$$

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

For a set of functions $\mathcal{F}$ and elements $x_1, \ldots, x_q \in \mathcal{M}(R)$, let $S_{\mathcal{F}}(x_1, \ldots, x_q)$ denote the cardinality of the set $\{(f(x_1), \ldots, f(x_q)) \mid f \in \mathcal{F}\}$. Let $\mathcal{F}_1$ denote the set 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

$$S_{\mathcal{F}_1}(X_1, \ldots, X_n, X_1', \ldots, X_n')$$

$$\leq \left|\left\{(\mathbb{1}_W(x_j^{(i)}))_{i=1,\ldots,2n,j=1,\ldots,n_i} \mid W = \bigcap_{j=1}^k H_j,\ H_j \text{ half-space}\right\}\right|$$

$$\leq \left(2\left(\sum_{i=1}^n n_i + n_i'\right)\right)^{c_0 k d \log(k)},$$

using [28, Theorem 1], and [39, 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

$$\left(\mathbb{P}_\sigma\left(\sup_{f \in \mathcal{F}} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i(f(X_i) - f(X_i'))}{\sqrt{(P_n'f + P_nf)/2}} \geq t\right)\right)$$

$$\leq \left( 2 \sum_{i=1}^{n} (n_i + n_i') \right)^{c_0 k d \log(k)} e^{-nt^2},$$

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

$$\mathbb{P} \left( \sup_{f \in \mathcal{F}} \frac{(P_n' - P_n)f}{\sqrt{(P_n'f + P_nf)/2}} \geq t \right) \leq (4nN_{max})^{c_0 k d \log(k)} e^{-nt^2},$$

that proves the second inequality of Lemma 22. The first inequality of Lemma 22 derives the same way from the second inequality of Lemma 24.

We turn to the third inequality of Lemma 22. Let $Z$ denote the process

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

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

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

so that $Z \leq \sqrt{\sum_{j=1}^{k} Z_j^2}$. According to the bounded differences inequality ([6, 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

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

where $\sigma_1, \ldots, \sigma_n$ are i.i.d Rademacher variables. Now assume that $X_1, \ldots, X_n$ is fixed and $j = 1$. For a set $\mathcal{F}$ of real-valued functions we denote by $\mathcal{N}(\mathcal{F}, \varepsilon, \|.\|)$ its $\varepsilon$-covering number with respect to the norm $\|.\|$. Denoting by $\Gamma_0$, $\Gamma_1$ and $\Gamma_2$ the following sets

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

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

$$\gamma_{(\mathbf{c}, t_1)}^{(0)}(X_i) - \gamma_{(\mathbf{c}', t_1')}^{(0)}(X_i) = \frac{1}{M} \int \left[ \gamma_{(c_1, t_1)}^{(1)}(u) \gamma_{\mathbf{c}}^{(2)}(u) - \gamma_{(c_1', t_1')}^{(1)}(u) \gamma_{\mathbf{c}'}^{(2)}(u) \right] X_i(du).$$

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

$$(\gamma_{(\mathbf{c}, t_1)}^{(0)}(X_i) - \gamma_{(\mathbf{c}', t_1')}^{(0)}(X_i))^2 \leq \left( \varepsilon + \frac{1}{M} \int |\gamma_{\mathbf{c}}^{(2)}(u) - \gamma_{\mathbf{c}'}^{(2)}(u)| X_i(du) \right)^2.$$

Thus,

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

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 [28, 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 $\varepsilon < 2$,

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

Using Dudley's entropy integral (see, e.g., [6, Corollary 13.2]) yields, for $k \geq 2$,

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

hence the result.

### A.1.4. 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.

$$\mathbb{E}\left(\left\|\left(\int(c_j-u)\mathbb{1}_{W_j(\mathbf{c})}(u)(\bar{X}_n-\mathbb{E}(X))(du)\right)_{j=1,\ldots,k}\right\|\right)$$

$$\leq \sqrt{\mathbb{E}\left\|\left(\int(c_j-u)\mathbb{1}_{W_j(\mathbf{c})}(u)(\bar{X}_n-\mathbb{E}(X))(du)\right)_{j=1,\ldots,k}\right\|^2}$$

$$\leq \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left(\left\|\left(\int(c_j-u)\mathbb{1}_{W_j(\mathbf{c})}(u)(X_i-\mathbb{E}(X))(du)\right)_{j=1,\ldots,k}\right\|^2\right)}$$

$$\leq \sqrt{\frac{(4RM)^2 k}{n}} = \frac{4RM\sqrt{k}}{\sqrt{n}}$$

A bounded difference inequality (see, e.g., [6, Theorem 6.2] entails that, with probability larger than $1 - e^{-x}$,

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

hence the result.

### A.1.5. Proof of Lemma 24

We follow the proof of [4, Theorem 1]. Let $t > 0$, and assume that $Pf - P_nf > 2t\sqrt{Pf}$, as well as $P'_nf \geq Pf - t\sqrt{Pf} \geq 0$. Let $g_a : \mathbb{R}^+ \to \mathbb{R}$ be defined as $g_a(x) = \frac{x-a}{\sqrt{x+a}}$, for $a \geq 0$. Then $g_a$ is nondecreasing on $\mathbb{R}^+$. With $a = P_nf$, $0 \leq x_2 = Pf - t\sqrt{Pf} \leq P'_nf = x_1$, we have $g_a(x_2) \leq g_a(x_1)$, so that

$$\frac{P'_nf - P_nf}{\sqrt{\frac{1}{2}(P'_nf + P_nf)}} \geq \frac{Pf - t\sqrt{Pf} - P_nf}{\sqrt{\frac{1}{2}(P_nf + Pf - t\sqrt{Pf})}}.$$

Since $P_nf + Pf - t\sqrt{Pf} \leq 2Pf$, we deduce that

$$\frac{P'_nf - P_nf}{\sqrt{\frac{1}{2}(P'_nf + P_nf)}} \geq \frac{P_f - P_nf - t\sqrt{Pf}}{\sqrt{Pf}} \geq t.$$

Thus,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\frac{P'_nf - P_nf}{\sqrt{\frac{1}{2}(P'_nf + P_nf)}}\right)$$

$$\geq \mathbb{P}\left(\exists f_0 \in \mathcal{F} \mid Pf_0 - P_nf_0 > 2t\sqrt{Pf_0} \text{ and } P'_nf_0 \geq Pf_0 - t\sqrt{Pf_0}\right).$$

Let $A = \left\{\exists f_0 \in \mathcal{F} \mid Pf_0 - P_nf_0 > 2t\sqrt{Pf_0}\right\}$. The previous inequality may be written as

$$\mathbb{P}\left(\exists f_0 \in \mathcal{F} \mid Pf_0 - P_n f_0 > 2t\sqrt{Pf_0} \text{ and } P'_n f_0 \geq Pf_0 - t\sqrt{Pf_0}\right)$$

$$\geq \mathbb{E}\left(\mathbb{1}_A \mathbb{P}\left(\bigcup_{\{f \mid Pf - P_n f > 2t\sqrt{Pf}\}} \{P'_n f \geq Pf - t\sqrt{Pf}\} \,\middle|\, X_1, \ldots, X_n\right)\right),$$

with, for a fixed $f \in \mathcal{F}$, using Cantelli's inequality,

$$\mathbb{P}\left(P'_n f - Pf < -2t\sqrt{Pf}\right) \leq \frac{\frac{1}{n}\operatorname{Var}(f)}{\frac{1}{n}\operatorname{Var}(f) + t^2 Pf}.$$

Since, for any $f \in \mathcal{F}$, $\|f\|_\infty \leq 1$, we have $\operatorname{Var}(f) \leq Pf^2 \leq Pf$, thus

$$\mathbb{P}\left(P'_n f - Pf < -t\sqrt{Pf}\right) \leq \frac{1}{2},$$

provided that $nt^2 \geq 1$. Thus

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{P'_n f - P_n f}{\sqrt{\frac{1}{2}(P'_n f + P_n f)}} \geq t\right) \geq \frac{1}{2}\mathbb{P}\left(\frac{Pf - P_n f}{\sqrt{Pf}} \geq 2t\right).$$

The other inequality proceeds from the same reasoning, considering $f$ such that $P_n f - Pf > 2t\sqrt{P_n f}$ and $P'_n f \leq Pf + t\sqrt{P_n f}$, and $g_a : \mathbb{R}^+ \to \mathbb{R}$ defined by $g_a(x) = \frac{a-x}{\sqrt{a+x}}$, that is nonincreasing for $a \geq 0$. Choosing $a = P_n f$, $0 \leq x_1 = P'_n f \leq Pf + t\sqrt{P_n f} = x_2$ leads to

$$\frac{P_n f - P'_n f}{\sqrt{\frac{1}{2}(P'_n f + P_n f)}} \geq \frac{P_n f - Pf - t\sqrt{P_n f}}{\sqrt{\frac{1}{2}\left(Pf + t\sqrt{P_n f} + P_n f\right)}} \geq \frac{P_n f - Pf - t\sqrt{P_n f}}{\sqrt{P_n f}} \geq t,$$

since $Pf + t\sqrt{P_n f} \leq P_n f$. Using Cantelli's inequality again leads to the result.

### A.2. Proof of Lemma 12

**Lemma** (12). *reach $\tau$. Let $S$ be a compact subset of $\mathbb{R}^d$, and $D$ denote the persistence diagram of the distance function $d_S$. 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.*

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

First, the definition of generalized gradient of $d_S$ – see [9] or [5, Section 9.2] – implies that the critical points of $d_S$ are all contained in the convex hull of $S$. As a consequence, they are all contained in the sublevel set $d_S^{-1}([0, 2\operatorname{diam}(S)])$. It follows from the Isotopy Lemma – [5, Theorem 9.5] – that all the sublevel sets $d_S^{-1}([0, t])$, $t > 2\operatorname{diam}(S)$ have the same homology. As a consequence, no point in $D$ has a larger coordinate than $2\operatorname{diam}(S)$ and $D$ is contained in $[0, 2\operatorname{diam}(S)]^2$.

Since $S$ is compact, the persistence module of the filtration defined by the sublevel sets of $d_S$ is $q$-tame ([10, Corollary 3.35]). 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 $[0, 2\mathrm{diam}(S)]^2$ 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.