



HAL
open science

A k-points-based distance for robust geometric inference

Claire Bréchet, Clément Levrard

► **To cite this version:**

Claire Bréchet, Clément Levrard. A k-points-based distance for robust geometric inference. 2019.
hal-02266408

HAL Id: hal-02266408

<https://hal.science/hal-02266408>

Preprint submitted on 14 Aug 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A k -points-based distance for robust geometric inference*

BréchetEAU, Claire
claire.brecheteau@ec-nantes.fr
École Centrale de Nantes – LMJL

Levrard, Clément
levrard@math.univ-parididerot.fr
Université Paris Diderot – LPMA

August 13, 2019

Abstract

Analyzing the sub-level sets of the distance to a compact sub-manifold of \mathbb{R}^d is a common method in topological data analysis, to understand its topology. Therefore, topological inference procedures usually rely on a distance estimate based on n sample points [41]. In the case where sample points are corrupted by noise, the distance-to-measure function (DTM, [16]) is a surrogate for the distance-to-compact-set function. In practice, computing the homology of its sub-level sets requires to compute the homology of unions of n balls ([28, 14]), that might become intractable whenever n is large.

To simultaneously face the two problems of a large number of points and noise, we introduce the k -power-distance-to-measure function (k -PDTM). This new approximation of the distance-to-measure may be thought of as a k -pointbased approximation of the DTM. Its sublevel sets consist in unions of k balls, and this distance is also proved robust to noise. We assess the quality of this approximation for k possibly drastically smaller than n , and provide an algorithm to compute this k -PDTM from a sample. Numerical experiments illustrate the good behavior of this k -points approximation in a noisy topological inference framework.

1 Introduction

Let $M \subset \mathbb{R}^d$ be a compact set whose topology is to be inferred from an n -sample $\mathbb{X}_n = \{X_1, X_2, \dots, X_n\}$ drawn on M . The pioneering work [37] has paved the way of topological inference, showing that the Devroye-Wise estimator $\bigcup_{i=1}^n \overline{B}(X_i, r)$, the union of closed Euclidean balls of radius r centered at sampled points, has the same homology as M , provided that M is a compact submanifold and r is well-chosen according to n . This result can be thought of as a particular instance of topological inference based on distance estimation: if d_M denotes the distance to M , then d_M is inferred via $d_{\mathbb{X}_n}$, the distance to sample points. The method exposed in [37] then boils down to infer the homology of the 0-sublevel set of d_M , $d_M^{-1}((-\infty, 0])$ from the homology of the r -sublevel set of $d_{\mathbb{X}_n}$. A general framework for geometric inference based on distance function estimation can be found in [16]. In a nutshell, [16, Proposition 4.3] states that if \hat{d} is an

*This work was partially supported by the ANR project TopData and GUDHI and Ecole Centrale de Nantes

estimator for d_M and M is smooth enough, then, for some $r > 0$, the r -sublevel set of \hat{d} has the same homology as M provided that $\|d_M - \hat{d}\|_\infty$ is small enough.

This distance estimation problem has been thoroughly investigated through the lens of Hausdorff set estimation: indeed, if \hat{M} is a set estimator, $\hat{d} = d_{\hat{M}}$, and d_H denotes the Hausdorff distance, then $d_H(\hat{M}, M) = \|\hat{d} - d_M\|_\infty$. Optimal rates of convergence for $\|\hat{d} - d_M\|_\infty$, in terms of sample size n have been derived under various types of regularity assumptions on M and noise conditions. In the noise-free case, optimal rates for $\|\hat{d} - d_M\|_\infty$ are given in [39] whenever M satisfies some convexity-type assumptions, whereas [27, 29, 2, 1] provide optimal rates when M is a smooth compact manifold. Note that, in the smooth manifold case with noisy observations, additional results on optimal rates for Hausdorff estimation can be found in [27]. All of these bounds can be combined with the aforementioned result [16, Proposition 4.3] to assess that the homology of M may be retrieved from the sublevel sets of \hat{d} , provided that n is large enough.

However, when the sample size n is large, computing the homology from an n -pointbased distance estimator \hat{d} may be computationally intractable. For instance, in the simplest case where $\hat{d} = d_{\mathbb{X}_n}$, a standard way to compute the homology of a sublevel set of $d_{\mathbb{X}_n}$ is to build a Rips complex based on \mathbb{X}_n whose homology can be efficiently computed [41]. The construction of such a simplicial complex requires the computation of pairwise distances, that is n^2 distances. To reduce this computational cost, a practical solution is to extract a *coreset* $\mathbb{X}_k \subset \mathbb{X}_n$ such that $\|d_{\mathbb{X}_k} - d_M\|_\infty$ is small enough to ensure topological correctness, then to compute a Rips complex based on \mathbb{X}_k . In the noise-free case, extracting such a coreset boils down to find an ε -covering of \mathbb{X}_n , where ε is the desired sup norm precision. Using a uniform grid shows that an ε -covering with $k(\varepsilon) \lesssim \varepsilon^{-\frac{1}{d}}$ points at most exists, and can be found in practice using farthest point sampling algorithm for instance [24]. Such a coreset may also be used to compute more involved estimates for d_M , as in [33].

In noisy settings, with observations of the type $X_i = Y_i + N_i$, Y_i on M and N_i denoting Gaussian noise, using a covering of sample points as base points for a coreset can lead to arbitrarily poor estimation of d_M . The goal of this paper is to nonetheless provide a coreset in such noisy situations, that is to build an approximation of d_M , based on the computation of a distance to k points, that may be proved close enough to d_M to allow further geometric inference.

To be more precise, we will build our k -points distance approximation as an approximation of the *distance-to-measure* [16], that may be thought of as a robust surrogate for d_M . Namely, for a Borel probability measure P on \mathbb{R}^d , a mass parameter $h \in [0, 1]$ and $x \in \mathbb{R}^d$, the distance of x to the measure P (DTM), $d_{P,h}(x)$ is defined by

$$d_{P,h}^2(x) = P_{x,h} \|x - \cdot\|^2,$$

where $P_{x,h}$ is the probability distribution defined as the restriction of the distribution P to the ball centered at x , with P -mass h , and with the notation Qf for the expectation of the function f with respect to the distribution Q . When P is uniform enough on M and M is regular enough, this distance is proved to approximate well the distance to M ([16, Proposition 4.9]) and is robust to noise ([16, Theorem 3.5]).

The distance-to-measure is usually inferred from \mathbb{X}_n via its empirical counterpart, called *empirical DTM*, replacing P by the empirical distribution $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$, where δ_x is the Dirac mass on x . As noted in [28], the sublevel sets of empirical DTM are union of around $\binom{n}{q}$ balls, with $q = hn$, which makes their computation intractable in practice. To bypass this issue, approximations of the empirical DTM have been proposed in [28] (q -witnessed distance) and [14] (power distance). Up to our knowledge, these are the only available approximations of the empirical DTM. The sublevel sets of these two approximations are union of n balls. Thus, it makes the computation of topological invariants more tractable for small data sets, from alpha-shapes for instance (see, e.g.,

[22]). Nonetheless, when n is large, there is still a need for an optimal set of points allowing to efficiently compute an approximation of the DTM, as pointed out in [38]. Up to our knowledge, the only results on such a reduction are on the negative side, exposing a lower bound on the number of points $k(\varepsilon)$ that are needed to build an ε -approximation of the empirical DTM [36].

The main contribution of this paper is the construction (Section 2.3), for a distribution P and a mass parameter h , of a k -power distance $d_{P,h,k}$ of the form

$$d_{P,h,k}(x) = \sqrt{\min_{i \in [1,k]} \|x - \tau_i\|^2 + \omega_{P,h}^2(\tau_i)},$$

that we call *k-power-distance-to-measure*, k -PDTM for short. We will prove that this k -points power distance is robust to noise (Proposition 17), and is a provably good approximation of the distance-to-measure (Proposition 14). This will allow us to give bounds on $\|d_{P,h,k} - d_M\|_\infty$ (Proposition 18) that can be used for further topological inference based on the sublevel sets of $d_{P,h,k}$. We then prove that its empirical counterpart $d_{P_n,h,k}$ is an optimal approximation of $d_{P,h,k}$ from an n -sample (Theorem 19 and Proposition 21). At last we provide a Lloyd's type algorithm [32] to compute in practice such a k -power distance based on n sample points (Section 3.3), and numerically illustrate its good performance in a framework of topological inference (Section 4).

The paper is organized as follows. Section 2 introduces definitions, notations and base results that are required for the construction of the k -PDTM. A proper definition of $d_{P,h,k}$ is given in Section 2.3, along with some basic properties. Section 3 exposes the main theoretical results of the paper, that consist in guarantees for the k -PDTM in a topological inference framework (Section 3.1), optimality of the sample approximation of the k -PDTM (Section 3.2), and an algorithm to compute it (Section 3.3). Numerical illustrations are given in Section 4, and Section 5, 6 gather the derivations of the main results. Proof of technical intermediate results as well as additional figures are deferred to the Appendix.

2 Notations, definitions and first results

2.1 Notations for the distance-to-measure

Throughout the paper, observations will be elements of the Euclidean space $(\mathbb{R}^d, \|\cdot\|)$. The ball centered at c with radius r is denoted by $B(c, r) = \{x \in \mathbb{R}^d \mid \|x - c\| < r\}$ and its closure by $\bar{B}(c, r)$. The sphere is denoted by $S(c, r) = \{x \in \mathbb{R}^d \mid \|x - c\| = r\}$. As well, if $A \subset \mathbb{R}^d$, \bar{A} denotes the closure of A , A° its interior, $\partial A = \bar{A} \setminus A^\circ$ its boundary and $A^c = \mathbb{R}^d \setminus A$ its complementary set in \mathbb{R}^d . For any positive integer k , $[1, k] = \{1, 2, \dots, k\}$. For any set A , $A^{(k)}$ stands for $\{\mathbf{t} = (t_1, t_2, \dots, t_k) \mid \forall i \in [1, k], t_i \in A\}$, where two elements are identified whenever they are equal up to a permutation of the coordinates. Following the quantization terminology, elements of $(\mathbb{R}^d)^{(k)}$ are called *codebooks* and their k elements *codepoints*. For any distribution P and any integrable function f , the integral of f with respect to P is denoted by Pf or $\int f(u)P(du)$. We also denote $\sup_x |f(x)|$ by $\|f\|_\infty$.

We consider probability distributions P with support $\text{Supp}(P) \subset \mathbb{R}^d$. The family of these distributions is denoted by $\mathcal{P}(\mathbb{R}^d)$. The subset of distributions P in $\mathcal{P}(\mathbb{R}^d)$ with finite moment of order 2 ($P\|\cdot\|^2 < \infty$) is denoted by $\mathcal{P}_2(\mathbb{R}^d)$. The distribution whose support is to be inferred is an element of $\mathcal{P}^K(\mathbb{R}^d) = \{P \in \mathcal{P}(\mathbb{R}^d) \mid \text{Supp}(P) \subset \bar{B}(0, K)\}$ for $K > 0$. To infer $\text{Supp}(P)$, we use a modified version Q of P . This measure Q is assumed to be *sub-Gaussian* with variance $V^2 > 0$. That is, Q is a distribution in $\mathcal{P}(\mathbb{R}^d)$

such that

$$Q(B(0, t)^c) \leq \exp\left(-\frac{t^2}{2V^2}\right)$$

for all $t > V$. The set of such measures is denoted by $\mathcal{P}^{(V)}(\mathbb{R}^d)$. Given $\mathbb{X}_n = \{X_1, X_2, \dots, X_n\}$ an n -sample from P , we denote by $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ the corresponding empirical distribution.

For $P \in \mathcal{P}_2(\mathbb{R}^d)$ and $h \in (0, 1]$, we use the notation $\mathcal{P}_h(P)$ for the set of distributions $P_h = \frac{1}{h}\mu$ with μ a submeasure of P (i.e. such that $\mu(B) \leq P(B)$ for every Borel set $B \subset \mathbb{R}^d$) satisfying $\mu(\mathbb{R}^d) = h$. The set of all of their expectations is defined by $\tilde{\mathcal{M}}_h(P) = \{m(P_h) \mid P_h \in \mathcal{P}_h(P)\}$, with the notation $m(P_h) = \int u P_h(du)$ for the mean of P_h , $v(P_h) = \int \|u - m(P_h)\|^2 P_h(du)$ for its variance and $M(P_h) = \|m(P_h)\|^2 + v(P_h)$ for its order 2 moment.

Some distributions in $\mathcal{P}_h(P)$ will be of special interest. Denote by $\delta_{P,h}(x) = \inf\{r > 0 \mid P(\bar{B}(x, r)) > h\}$ the smallest radius of a ball centered at $x \in \mathbb{R}^d$ of P -mass h . Then, *local distributions* are defined as restrictions of P to these balls.

Definition 1.

Let $P \in \mathcal{P}(\mathbb{R}^d)$. The set of local distributions at a point x , with mass parameter $h \in (0, 1]$, denoted by $\mathcal{P}_{x,h}(P)$ is the set of distributions $P_{x,h}$ defined by $P_{x,h} = \frac{1}{h}\mu$, where μ satisfies:

1. μ is a submeasure of P with P -mass h : $\frac{1}{h}\mu \in \mathcal{P}_h(P)$.
2. μ coincides with P on $B(x, \delta_{P,h}(x))$.
3. $\text{Supp}(\mu) \subset \bar{B}(x, \delta_{P,h}(x))$.

Note that when $P(\partial B(x, \delta_{P,h}(x))) = 0$, the set of distributions $\mathcal{P}_{x,h}(P)$ is reduced to a singleton $\{P_{x,h}\}$ with $P_{x,h}(B) = \frac{1}{h}P(B \cap B(x, \delta_{P,h}(x)))$, for any Borel set B . Accordingly, we may define a *local mean* as the expectation of a local distribution, $m(P_{x,h}) = \int u P_{x,h}(du)$ associated to some $x \in \mathbb{R}^d$ and $h \in (0, 1]$. The set of local means of P with parameter $h \in (0, 1]$ is defined by $\mathcal{M}_h(P) = \{m(P_{x,h}) \mid x \in \mathbb{R}^d, P_{x,h} \in \mathcal{P}_{x,h}(P)\}$.

Example 2 (Uniform distribution on a circle). Let P denote the uniform distribution on the unit sphere $S(0, 1) \subset \mathbb{R}^2$. Then, for every $x \neq 0$, $\mathcal{P}_{x,h}(P)$ is the singleton $\{P_{x,h}\}$, $P_{x,h}$ being the uniform distribution on the arc centered at $\frac{x}{\|x\|}$ subtending an angle $2\pi h$. For $x = 0$, $\mathcal{P}_{0,h}(P)$ coincides with $\mathcal{P}_h(P)$. As a consequence, for $x \neq 0$, $m(P_{x,h}) = \text{sinc}(h\pi) \frac{x}{\|x\|}$ and $v(P_{x,h}) = 1 - \text{sinc}(h\pi)^2$ where $\text{sinc} : x \mapsto \frac{\sin(x)}{x}$ is the sinus cardinal function. For $x = 0$, the set $\{m(P_{0,h}) \mid P_{0,h} \in \mathcal{P}_{0,h}(P)\}$ coincides with the ball $\bar{B}(0, \text{sinc}(h\pi))$ and $v(P_{0,h}) \geq 1 - \text{sinc}(h\pi)^2$ with equality if and only if $P_{0,h} = P_{x,h}$ for some $x \neq 0$. Note that for such an example, $\tilde{\mathcal{M}}_h(P)$ coincides with the set of local means $\mathcal{M}_h(P)$.

These notions of local distributions and local means are required to define the notion of distance-to-measure.

2.2 Definition of the distance-to-measure (DTM)

In the framework of geometric inference, to face the non-robustness to noise of the function distance to a compact set, the notion of *distance-to-measure (DTM)* has been introduced in [16]. The DTM $d_{P,h}$ is a function defined on \mathbb{R}^d , associated with a probability distribution P and depending on a mass parameter $h \in [0, 1]$. A definition of

the DTM in terms of submeasures is given in [16, Proposition 3.3]. For every $x \in \mathbb{R}^d$, $h \in (0, 1]$ and $P_{x,h} \in \mathcal{P}_{x,h}(P)$,

$$d_{P,h}^2(x) = \inf_{P_h \in \mathcal{P}_h(P)} P_h \|x - \cdot\|^2 = P_{x,h} \|x - \cdot\|^2. \quad (1)$$

Whenever h is small, the DTM provably approximates well the distance to the support of P [16, Corollary 4.8 and Proposition 4.9], when P is a uniform distribution on a submanifold for instance.

If a small h allows to approximate the distance to the support of P , larger values for h make the DTM robust to small variations of the distribution P , in terms of the Wasserstein metric. Indeed, according to [16], if P and Q are two probability distributions on the space $(\mathbb{R}^d, \|\cdot\|)$ with finite second moment, then

$$\|d_{P,h} - d_{Q,h}\|_\infty \leq \frac{1}{\sqrt{h}} W_2(P, Q). \quad (2)$$

Let us recall that the Wasserstein metric W_p is defined, for $p \geq 1$, by

$$W_p^p(P, Q) = \inf_{\pi \in \Pi(P, Q)} \mathbb{E}_{(X, Y) \sim \pi} [\|X - Y\|^p], \quad (3)$$

where $\Pi(P, Q)$ denotes the set of distributions on $\mathbb{R}^d \times \mathbb{R}^d$ of random vectors (X, Y) such that $X \sim P$ (i.e. X is a random variable of distribution P) and $Y \sim Q$.

According to (1), the distance-to-measure can be written as a *power distance*, that is the square root of a function $f_{\tau, \omega} : x \mapsto \inf_{i \in I} \|x - \tau_i\|^2 + \omega_i^2$ for some set I , a family of centers $\tau = (\tau_i)_{i \in I}$ and weights $\omega = (\omega_i)_{i \in I}$. Special instances of power distances are *k-power distances*, indexed on a finite set of cardinal $|I| = k$. Concerning the distance-to-measure, for every $P \in \mathcal{P}_2(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$, we have

$$d_{P,h}^2(x) = \inf_{P_h \in \mathcal{P}_h(P)} \|x - m(P_h)\|^2 + v(P_h), \quad (4)$$

where the minimum is attained at any measure $P_h = P_{x,h}$ in $\mathcal{P}_{x,h}(P)$. In this case, the centers $m(P_h)$ are elements of $\tilde{\mathcal{M}}_h(P)$. Note that according to the second equality of Equation (1), $\mathcal{P}_h(P)$ can be replaced by $\bigcup_{x \in \mathbb{R}^d} \mathcal{P}_{x,h}(P)$ in (4), so that the centers are actually elements of $\mathcal{M}_h(P)$. The following example gives some intuition on why the distance-to-measure can be a convenient tool for geometric inference in noisy settings, compared to classical quantization-based approaches.

Example 3 (Uniform distribution on a circle with noise). *Let $Q_\beta = \beta \mathcal{U}_{S(0,1)} + (1 - \beta) \mathcal{U}_{B(0,1)}$ be a noisy version of $P = \mathcal{U}_{S(0,1)}$, the uniform distribution on the circle, for some $\beta \in (0, 1)$. According to [16, Theorem 3.5, Corollary 4.8], since $W_2(Q_\beta, P) \leq \sqrt{1 - \beta}$, we have $\|d_{Q_\beta, h} - d_{S(0,1)}\|_\infty \leq Ch + \sqrt{\frac{1 - \beta}{h}}$ for some $C > 0$. Thus, for $h > 81(1 - \beta)$ and $1 - \beta$ small enough, [16, Theorem 4.6] ensures that the r -sublevel sets of $d_{Q_\beta, h}$ are homotopy equivalent to $S(0, 1)$, for a range of r 's.*

Now let τ^ be a minimizer of the k -means criterion $Q_\beta \min_{j \in \llbracket 1, k \rrbracket} \|\cdot - \tau_j\|^2$, in other words, an optimal k -points codebook for Q_β . The following Lemma shed some light on the approximation properties of the distance to τ^* function.*

Lemma 4.

Let d_{τ^} denote the distance to τ^* function. Then, for k large enough,*

$$\sup_{x \in S(0,1)} |d_{\tau^*}(x) - d_{S(0,1)}(x)| \leq C \left(\frac{1}{k^2} + (1 - \beta) \right)^{\frac{1}{3}}.$$

for some constant $C > 0$. On the other hand, for every $\rho > 0$, there exists $k_{\rho, \beta}$ such that, for all $k \geq k_{\rho, \beta}$, τ^ has at least one codepoint in $B(0, \rho)$.*

A direct consequence of Lemma 4 is

$$\sup_{k \geq 0} \|d_{S(0,1)} - d_{\tau^*}\|_{\infty} \geq \sup_{k \geq 0} |d_{S(0,1)}(0) - d_{\tau^*}(0)| = 1.$$

A proof of Lemma 4 is given in Section A.1 of the Appendix. The intuition behind Lemma 4 is that though optimal codebooks designed via classical quantization can yield provably good covering of topological structures such as manifolds, they are also likely to have codepoints far from the structure in some noisy cases. In this case, geometric inference based on the sublevel sets of d_{τ^*} might lead to poor results.

2.3 Definition of the k -PDTM

As illustrated above, in Example 3, the distance-to-measure may be thought of as a robustified version of the distance-to-compact-set, designed for geometric inference in noisy settings. According to (4), its sub-level sets are unions of balls centered at elements of $\tilde{\mathcal{M}}_h(P)$. As noted in [28], in general, for empirical distributions based on n points $\{X_1, \dots, X_n\}$, this amount of balls is finite but may be large (of order $\binom{n}{nh}$, where h is the mass parameter of the DTM). Approximations of the DTM consisting in reducing this number of balls to the sample size n are exposed in [28, 14]. In this paper, we propose to reduce this number of balls to some $k \in \mathbb{N}^*$ possibly much smaller than the sample size, resulting in an approximation of the distance-to-measure that we prove accurate enough for further topological inference. This section is devoted to the introduction of such an approximation, namely the k -PDTM.

The k -PDTM is an approximation of the DTM obtained after reducing the set of submeasures $\mathcal{P}_h(P)$ (or equivalently, the set of centers $\tilde{\mathcal{M}}_h(P)$) to a set of k well-chosen submeasures (or k centers) in the definition of the DTM (4). As an answer to [38], such a set of k centers may be considered as a coresets for the DTM. These k submeasures are obtained by minimizing the following criterion R .

Definition 5.

For $P \in \mathcal{P}_2(\mathbb{R}^d)$ and $\mathbf{P} = (P_i)_{i \in [1,k]} \in \mathcal{P}_h(P)^{(k)}$, we define $R(\mathbf{P})$ by

$$R(\mathbf{P}) = P \min_{i \in [1,k]} \|\cdot - m(P_i)\|^2 + v(P_i).$$

The following Proposition 6 ensures that there exist optimal submeasures with respect to the risk R .

Proposition 6.

If $P \in \mathcal{P}_2(\mathbb{R}^d)$, then the minimum of R is attained in $\mathcal{P}_h(P)^{(k)}$. We denote by \mathbf{P}^* any such minimizer.

The proof of Proposition 6 is to be found in Section 5.1. These optimal submeasures allow us to define the k -PDTM as follows.

Definition 7.

The k -PDTM is any function $d_{P,h,k} : \mathbb{R}^d \mapsto \mathbb{R}$ defined by

$$d_{P,h,k}^2(x) = \min_{i \in [1,k]} \|x - m(P_i^*)\|^2 + v(P_i^*),$$

for some $\mathbf{P}^* = (P_1^*, \dots, P_k^*) \in \arg \min_{\mathbf{P} \in \mathcal{P}_h(P)^{(k)}} R(\mathbf{P})$.

The k -PDTM is a k -power distance which graph lies above the graph of the DTM. It is not necessarily uniquely defined, since several minimizers of R may exist. Its sublevel sets are unions of k balls. Besides, the k centers of the k -PDTM yield a decomposition of the space \mathbb{R}^d into k cells, and consequently, a decomposition of P into k weighted Voronoi measures.

Definition 8.

A set of weighted Voronoi measures associated to a distribution $P \in \mathcal{P}_2(\mathbb{R}^d)$, k submeasures $(P_i)_{i \in \llbracket 1, k \rrbracket} \in \mathcal{P}_h(P)^{(k)}$ and $h \in (0, 1]$ is a collection $\{\tilde{P}_{1,h}, \tilde{P}_{2,h}, \dots, \tilde{P}_{k,h}\}$ of $k \in \mathbb{N}^*$ non-negative submeasures of P such that $\sum_{i=1}^k \tilde{P}_{i,h} = P$ and

$$\forall x \in \text{Supp}(\tilde{P}_{i,h}), \|x - m(P_i)\|^2 + v(P_i) \leq \|x - m(P_j)\|^2 + v(P_j), \forall j \in \llbracket 1, k \rrbracket.$$

Note that a set of weighted Voronoi measures can always be assigned to any $P \in \mathcal{P}_2(\mathbb{R}^d)$ and $(P_i)_{i \in \llbracket 1, k \rrbracket} \in \mathcal{P}_h(P)^{(k)}$. Indeed, \mathbb{R}^d may be splitted into weighted Voronoi cells associated to the centers $(m(P_i))_{i \in \llbracket 1, k \rrbracket}$ and weights $(v(P_i))_{i \in \llbracket 1, k \rrbracket}$ ([9, Section 4.4.2]), with ties arbitrarily broken. The following key property of weighted Voronoi measures implies that minimizers \mathbf{P}^* of the criterion R are actually elements of $(\bigcup_{t \in \mathbb{R}^d} \mathcal{P}_{t,h}(P))^{(k)}$. With a slight abuse of notation, for every $\mathbf{t} \in (\mathbb{R}^d)^{(k)}$, we denote the quantity $R(P_{t_1,h}, \dots, P_{t_k,h})$ by $R(\mathbf{t})$, where $P_{t_i,h} \in \mathcal{P}_{t_i,h}(P)$.

Proposition 9.

Let $P \in \mathcal{P}_2(\mathbb{R}^d)$, and $(P_i)_{i \in \llbracket 1, k \rrbracket} \in \mathcal{P}_h(P)^{(k)}$. Then

$$R(m(\tilde{P}_{1,h}), \dots, m(\tilde{P}_{k,h})) \leq R(P_1, \dots, P_k),$$

with equality if and only if, for all $i \in \llbracket 1, k \rrbracket$ such that $\tilde{P}_{i,h}(\mathbb{R}^d) \neq 0$, we have $P_i \in \mathcal{P}_{m(\tilde{P}_{i,h}),h}(P)$.

The proof of Proposition 9 is deferred to Section 5.2. This proposition provides a natural and tractable procedure for computing local optima of the criterion R , cf. Algorithm 1 in Section 3.3. Indeed, Proposition 9 shows that minimizing R over the set of k submeasures boils down to minimize $\mathbf{t} \mapsto R(\mathbf{t})$ over $(\mathbb{R}^d)^{(k)}$. An alternative definition of the k -PDTM in terms of local distributions may be stated accordingly.

Corollary 10.

The k -PDTM is any function $d_{P,h,k} : \mathbb{R}^d \mapsto \mathbb{R}$ defined by

$$d_{P,h,k}^2(x) = \min_{i \in \llbracket 1, k \rrbracket} \|x - m(P_{t_i^*,h})\|^2 + v(P_{t_i^*,h})$$

for some $\mathbf{t}^* = (t_1^*, \dots, t_k^*) \in \arg \min_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} R(\mathbf{t})$.

This definition of the k -PDTM in terms of local means and variances is very convenient for the purpose of its computation, but the more general definition of the k -PDTM in terms of submeasures (Definition 7) is also crucial. Indeed, from Definition 7 we may also state an alternative parametrization of the k -PDTM by elements $\tau = m(P_h)$ of $\tilde{\mathcal{M}}_h(P)$ that will allow for a geometric interpretation of the k -PDTM. The corresponding variances $v(P_h)$ will be obtained as images $\omega_{P,h}(\tau)$ of the function $\omega_{P,h}$ defined for every τ in \mathbb{R}^d by

$$\omega_{P,h}^2(\tau) = \sup_{x \in \mathbb{R}^d} d_{P,h}^2(x) - \|x - \tau\|^2.$$

Lemma 11.

If $P \in \mathcal{P}_2(\mathbb{R}^d)$, then $\omega_{P,h}(\tau) < +\infty$ if and only if $\tau \in \tilde{\mathcal{M}}_h(P)$. Moreover, if $\tau \in \tilde{\mathcal{M}}_h(P)$, then there exists $P_h \in \mathcal{P}_h(P)$ such that $\tau = m(P_h)$ and $\omega_{P,h}^2(\tau) = v(P_h)$. More precisely, $\omega_{P,h}^2(\tau) = \inf_{P_h \in \mathcal{P}_h(P), m(P_h) = \tau} v(P_h)$.

The proof of Lemma 11 is deferred to Section 5.3. The natural reparametrization of the criterion R with the set of centers follows.

Theorem 12.

Let $P \in \mathcal{P}_2(\mathbb{R}^d)$. Then $\mathbf{t}^* \in \arg \min_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} R(\mathbf{t})$ if and only if

$$\left(m(P_{t_1^*, h}), \dots, m(P_{t_k^*, h})\right) \in \arg \min_{\tau \in (\mathbb{R}^d)^{(k)}} P \min_{i \in [1, k]} \|\cdot - \tau_i\|^2 + \omega_{P, h}^2(\tau_i).$$

The proof of Theorem 12 is to be found in Section 5.4. Theorem 12 states that the k -PDTM is a solution of a weighted k -meantype criterion. According to Lemma 11, the regularization terms $\omega_{P, h}(\tau_i)$ force the optimal codebooks τ^* to be in $\tilde{\mathcal{M}}_h(P)^{(k)}$. Intuitively, elements τ such that $\omega_{P, h}(\tau)$ is small will be favoured. Such τ 's gather a proportion h of the mass of P on their neighborhood. On the contrary, for elements τ such that $\omega_{P, h}(\tau)$ is large, the corresponding weighted Voronoi measures will not be massive, and the ball associated to such τ 's will appear in the r -sublevel set of the function $x \mapsto \|\cdot - \tau\|^2 + \omega_{P, h}^2(\tau)$ for large r 's only. A direct consequence of Theorem 12 is that the squared k -PDTM may be interpreted as the closest squared k -power distance to the squared DTM from above, in terms of $L_1(P)$ norm. This interpretation comes from the straightforward inequality $\|x - \tau\|^2 + \omega_{P, h}^2(\tau) \geq d_{P, h}^2(x)$. The resulting inequality $d_{P, h, k}^2 \geq d_{P, h}^2$ allows for further comparison with k -means approximation of the distance-to-compact-set in noisy settings.

Example 13 (Noisy distribution on the circle). *For the distribution $Q_\beta = \beta \mathcal{U}_{S(0,1)} + (1 - \beta) \mathcal{U}_{B(0,1)}$. If $h > 1 - \beta$, since $d_{Q_\beta, h, k}^2(0) \geq d_{Q_\beta, h}^2(0)$, we have $d_{Q_\beta, h, k}^2(0) \geq 1 - \frac{1-\beta}{h}$. As a consequence, $\inf_{k \geq 0} d_{Q_\beta, h, k}^2(0) \geq 1 - \frac{1-\beta}{h}$, whereas $\inf_{k \geq 0} d_{\tau^*}^2(0) = 0$, where τ^* denotes an optimal k -points codebook.*

The above example shows that we can expect the k -PDTM to approximate well the distance-to-compact-set in remote areas, contrary to the distance based on k -means, d_{τ^*} . To check whether the k -PDTM provides also an efficient covering of the targeted structure is investigated in the following section.

3 Theoretical results for the k -PDTM

3.1 Geometric inference with the k -PDTM

Let M be a compact subset of \mathbb{R}^d , and let $f_M(\varepsilon)$ denote the ε -covering number of M (minimum number of balls of radius ε needed to cover M). Here we show that the k -PDTM approximates the DTM, provided that the covering number of M and the continuity modulus of the map $x \mapsto m(P_{x, h})$ are not too large. In what follows, C_{l_1, \dots, l_s} and c_{l_1, \dots, l_s} denote quantities depending on l_1, \dots, l_s only.

Proposition 14.

Let $P \in \mathcal{P}^K(\mathbb{R}^d)$ for $K > 0$ and let $M \subset B(0, K)$ be such that $P(M) = 1$. Let $f_M(\varepsilon)$ denote the ε -covering number of M . Then we have

$$0 \leq P d_{P, h, k}^2 - d_{P, h}^2 \leq 2 f_M^{-1}(k) \zeta_{P, h}(f_M^{-1}(k)), \quad \text{with} \quad f_M^{-1}(k) = \inf \{\varepsilon > 0 \mid f_M(\varepsilon) \leq k\},$$

where $\zeta_{P, h}$ is the continuity modulus of $x \mapsto m(P_{x, h})$, that is

$$\zeta_{P, h}(\varepsilon) = \sup_{x, y \in M, \|x - y\| \leq \varepsilon} \sup_{P_{x, h} \in \mathcal{P}_{x, h}(P), P_{y, h} \in \mathcal{P}_{y, h}(P)} \{|m(P_{x, h}) - m(P_{y, h})|\}.$$

A proof of Proposition 14 is given in Section 6.2. Whenever P is roughly uniform on its support, the quantities $f_M^{-1}(k)$ and $\zeta_{P, h}$ mostly depend on the dimension and radius of M . We illustrate this point with two instances of particular interest for geometric inference. First, the case where the distribution P has an ambient-dimensional support is investigated.

Corollary 15.

Assume that P has a density f satisfying $0 < f_{\min} \leq f \leq f_{\max}$. Then

$$0 \leq Pd_{P,h,k}^2 - d_{P,h}^2 \leq C_{f_{\max},K,d,h} k^{-2/d}.$$

The proof of Corollary 15 is given in Section 6.3. Note that no assumptions on the geometric regularity of M are required for Corollary 15 to hold. In the case where M has a lower-dimensional structure, more regularity is required, as for instance in the following corollary.

Corollary 16.

Suppose that P is supported on $N \subset B(0, K)$, a compact d' -dimensional \mathcal{C}^2 -submanifold. Assume that P has a density $0 < f_{\min} \leq f \leq f_{\max}$ with respect to the volume measure on N . Moreover, suppose that P satisfies, for all $x \in N$ and positive r ,

$$P(B(x, r)) \geq cf_{\min} r^{d'} \wedge 1. \quad (5)$$

Then, for $k \geq c_{N,f_{\min}}$ and $h \leq C_{N,f_{\min}}$, we have $0 \leq Pd_{P,h,k}^2 - d_{P,h}^2 \leq C_{N,f_{\min},f_{\max}} k^{-2/d'}$.

Note that (5), also known as (cf_{\min}, d') -standard assumption, is a usual assumption in the set estimation framework (see, e.g., [18]). In the submanifold case, it may be thought of as a condition preventing the boundary from being arbitrarily narrow. This assumption is satisfied for instance in the case where ∂N is empty or is a \mathcal{C}^2 $(d' - 1)$ -dimensional submanifold (see, e.g., [3, Corollary 1]). An important feature of Corollary 16 is that this approximation bound does not depend on the ambient dimension. The proof of Corollary 16 may be found in Section 6.4. Next we assess that our k -PDTM shares with the DTM the key property of robustness to noise.

Proposition 17.

Let $P \in \mathcal{P}^K(\mathbb{R}^d)$ for some $K > 0$ and $Q \in \mathcal{P}_2(\mathbb{R}^d)$. Let $d_{Q,h,k}^2$ denote a k -PDTM for Q . Then, $P \left| d_{Q,h,k}^2 - d_{P,h}^2 \right| \leq B_{P,Q,h,k}$, where

$$B_{P,Q,h,k} = 3 \|d_{Q,h}^2 - d_{P,h}^2\|_{\infty, B(0,K)} + Pd_{P,h,k}^2 - d_{P,h}^2 + 4W_1(P, Q) \sup_{s \in \mathbb{R}^d} \|m(P_{s,h})\|.$$

The proof of Proposition 17 can be found in Section 6.5. Note that Lemma 23 provides a bound on $\|m(P_{s,h})\|$ whenever P is sub-Gaussian. Moreover, [16, Theorem 3.5] ensures that $\|d_{Q,h}^2 - d_{P,h}^2\|_{\infty, B(0,K)}$ can be bounded in terms of $W_2(P, Q)$, up to a constant dependent on K . Note at last that bounds on $Pd_{P,h,k}^2 - d_{P,h}^2$ may be derived using Proposition 14.

Proposition 17 can provide guarantees on $Pd_{Q,h,k}^2$. In turn, provided that M is regular enough, these bounds can be turned into L_∞ bounds between $d_{Q,h,k}$ and d_M . Following [16, Section 4], these L_∞ bounds can guarantee that the sublevels sets of the k -PDTM are homotopy equivalent to M , under suitable assumptions.

Proposition 18.

Let M be a compact set in $B(0, K)$ such that $P(M) = 1$. Moreover, assume that there exists d' such that, for every $p \in M$ and $r \geq 0$,

$$P(B(p, r)) \geq C(P)r^{d'} \wedge 1. \quad (6)$$

Let Q be a Borel probability measure (thought of as a perturbation of P), and let Δ_P^2 denote $Pd_{P,h,k}^2$. Then, we have

$$\|d_{Q,h,k} - d_M\|_\infty \leq \max \left\{ C(P)^{-\frac{1}{d'+2}} \Delta_P^{\frac{2}{d'+2}}, 2\Delta_P, W_2(P, Q)h^{-\frac{1}{2}} \right\},$$

where W_2 denotes the Wasserstein distance.

The proof of Proposition 18 can be found in Section 6.6. According to [16, Corollary 4.8], if P satisfies (6), then

$$\|d_{Q,h} - d_M\|_\infty \leq \left(\frac{h}{C(P)}\right)^{\frac{1}{d'}} + W_2(P, Q)h^{-\frac{1}{2}}.$$

Hence, Proposition 18 ensures that the k -PDTM achieves roughly the same performance as the distance-to-measure provided that $d_{Q,h,k}^2$ is small enough on the support M to be inferred. As will be shown in the following section, this will be the case if Q is an empirical measure drawn close to the targeted support.

3.2 Approximation of the k -PDTM from point clouds

In this section, $P \in \mathcal{P}^K(\mathbb{R}^d)$ is a distribution supported on a compact set M to be inferred. We have at our disposal an n -sample $\mathbb{X}_n = \{X_1, X_2, \dots, X_n\}$ from a modification $Q \in \mathcal{P}_2(\mathbb{R}^d)$ of P . An approximation of the k -PDTM $d_{Q,h,k}$, is given by the empirical k -PDTM $d_{Q_n,h,k}$, where $Q_n = \sum_{i=1}^n \frac{1}{n} \delta_{X_i}$ is the empirical measure from \mathbb{X}_n . Note that when $k = n$, the empirical k -PDTM $d_{Q_n,h,n}$ coincides with the q -witnessed distance [28], for $q = nh$. We investigate the quality of approximation of the DTM $d_{P,h}$ with the empirical k -PDTM $d_{Q_n,h,k}$, when Q is defined as the convolution of P with a sub-Gaussian distribution with variance σ^2 . Within this context, according to Lemma 24, Q is sub-Gaussian with variance $V^2 = (K + \sigma)^2$.

Theorem 19.

Let P be supported on $M \subset B(0, K)$. Assume that we observe X_1, \dots, X_n such that $X_i = Y_i + Z_i$, where the Y_i 's and Z_i 's are all independent, Y_i is sampled from P and Z_i is sub-Gaussian with variance σ^2 , with $\sigma \leq K$. Let Q_n denote the empirical distribution associated with the X_i 's. Then, for any $p > 0$, with probability larger than $1 - 10n^{-p}$, we have

$$|P(d_{Q_n,h,k}^2 - d_{Q,h,k}^2)| \leq C\sqrt{kd} \frac{K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C \frac{K\sigma}{\sqrt{h}}.$$

A proof of Theorem 19 is given in Section 6.7. The \sqrt{kd}/\sqrt{n} term is in line with the rate of convergence for the k -means method (see, e.g., [8]), as well as with the rate of convergence for $\|d_{P_n,h} - d_{P,h}\|_\infty$ exposed in [17]. The $K\sigma$ term is due to the expectation with respect to P (instead of Q). Theorem 19, combined with Proposition 17, allows us to choose k in order to minimize $|Pd_{Q_n,h,k}^2 - d_{P,h}^2|$. Indeed, in the framework of Corollaries 15 and 16 where the support has intrinsic dimension d' , such a minimization boils down to optimizing a quantity of the form $\frac{C\sqrt{k}K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C_{P,h}k^{-\frac{2}{d'}}$. Choosing $k \sim n^{\frac{d'}{d'+4}}$ achieves the desired tradeoff between bias and variance. From the point of view of geometric inference, this leads to computing the distance to $n^{d'/(d'+4)}$ points rather than n , which might save some time. Note that when d' is large, smaller choices of k , though suboptimal for our bounds, would nonetheless give the right topology for large n . In some sense, Theorem 19 advocates only an upper bound on k , above which no increase of precision can be expected. Combining Theorem 19 and Proposition 14 leads to the following result.

Proposition 20.

With the same setting as Theorem 19, if M is a submanifold with intrinsic dimension $d' \geq 1$, then:

$$|Pd_{Q_n,h,k}^2 - d_{P,h}^2| \leq C\sqrt{kd} \frac{K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C \frac{K\sigma}{h} + C_{P,h}k^{-\frac{2}{d'}}.$$

Thus, choosing $k \sim n^{\frac{d'}{d'+4}}$ leads to

$$|Pd_{Q_n, h, k}^2 - d_{P, h}^2| \leq C_{P, h} \sqrt{dn}^{-\frac{2}{d'+4}} \frac{K^2((p+1)\log(n))^{\frac{3}{2}}}{h} + C \frac{K\sigma}{h}.$$

The proof of Proposition 20 is to be found in Section 6.8. Noting that $Pd_{P, h}^2 \leq C_P h^{\frac{1}{d'}}$ in this case (see, e.g., [16, Proposition 4.9]), Proposition 20 can be combined with Proposition 18 to yield a bound on $\|d_{Q_n, h, k} - d_M\|_\infty$.

To assess optimality of Theorem 19 in terms of sample size dependency, a lower bound on the best k -points approximation of the DTM that is achievable on the set of distributions supported in $B(0, K)$ may be derived from [7, Theorem 1] or [31, Proposition 3.1].

Proposition 21.

For $\mathbf{t} \in (\mathbb{R}^d)^{(k)}$ and P a probability measure, denote

$$d_{P, h, \mathbf{t}}^2 : x \mapsto \min_{j \in \llbracket 1, k \rrbracket} [\|x - m(P_{t_j, h})\|^2 + v(P_{t_j, h})].$$

For $k \geq 3$, $n \geq \frac{3k}{2}$ and $h \leq \frac{1}{2k}$, we have

$$\inf_{\hat{\mathbf{t}}} \sup_{P|\text{Supp}(P) \subset B(0, K)} \mathbb{E}P(d_{P, h, \hat{\mathbf{t}}}^2 - d_{P, h, k}^2) \geq c_0 \frac{K^2 k^{\frac{1}{2} - \frac{2}{d}}}{\sqrt{n}}, \quad (7)$$

where c_0 is a constant and $\hat{\mathbf{t}}$ denotes an empirically designed vector $(\hat{t}_1, \dots, \hat{t}_k)$ in $(\mathbb{R}^d)^{(k)}$. Moreover, if $n \geq 14k$, then

$$\inf_{\hat{\mathbf{t}}} \sup_{P|\text{Supp}(P) \subset B(0, K)} \mathbb{E}P(d_{P_n, h, \hat{\mathbf{t}}}^2 - d_{P_n, h, k}^2) \geq c_0 \frac{K^2 k^{\frac{1}{2} - \frac{2}{d}}}{\sqrt{n}} - 32K^2 k e^{-\frac{n}{72k^2}}. \quad (8)$$

Thus, Proposition 21 confirms that the sample size dependency of Theorem 19 is optimal in the noise-free case, up to $\log(n)$ factors. A proof is given in Section 6.9.

3.3 Algorithm

In this section we expose a Lloyd-type algorithm to compute a local minimizer for the cost function associated with the empirical k -PDTM. For an n -sample \mathbb{X}_n with empirical distribution Q_n , Proposition 9 suggests a procedure to minimize the empirical risk $\mathbf{t} \mapsto R_n(\mathbf{t}) = Q_n \min_{i \in \llbracket 1, k \rrbracket} \|\cdot - m(Q_n t_i, h)\|^2 + v(Q_n t_i, h)$. Indeed, given some codebook \mathbf{t} , replacing \mathbf{t} with the means of the weighted Voronoi measures $(\tilde{Q}_n t_i, h)_{i \in \llbracket 1, k \rrbracket}$ can only decrease the empirical risk R_n . For a sample \mathbb{X}_n , this boils down to compute the weighted Voronoi cells $(\mathcal{C}(t_i))_{i \in \llbracket 1, k \rrbracket}$ (i.e. the support of the measures $(\tilde{Q}_n t_i, h)_{i \in \llbracket 1, k \rrbracket}$), and to replace t_i with the mean of the points of \mathbb{X}_n in $\mathcal{C}(t_i)$. We use the notation $|\mathcal{C}(t)|$ for the cardinal number of $\mathcal{C}(t)$, $m(t)$ for $m(Q_n t, h)$ and $v(t)$ for $v(Q_n t, h)$. The procedure is described in Algorithm 1.

Algorithm 1: Local minimum algorithm

<p>Input : \mathbb{X}_n an n-sample from Q, h and k ; # Initialization Sample t_1, t_2, \dots, t_k from \mathbb{X}_n without replacement. ; while the t_is vary make the following two steps : # Decomposition into weighted Voronoi cells.</p>

```

for j in 1..n:
    Add  $X_j$  to the  $\mathcal{C}(t_i)$  (for  $i$  as small as possible) satisfying
     $\|X_j - m(t_i)\|^2 + v(t_i) \leq \|X_j - m(t_l)\|^2 + v(t_l) \forall l \neq i$ ;
    # Computation of the new centers.
    for i in 1..k:
         $t_i = \frac{1}{|\mathcal{C}(t_i)|} \sum_{X \in \mathcal{C}(t_i)} X$ ;
Output:  $(t_1, t_2, \dots, t_k)$ 

```

Proposition 22.

Algorithm 1 converges to a local minimum of

$$R_n : \mathbf{t} \mapsto Q_n \min_{i \in \llbracket 1, k \rrbracket} \|\cdot - m(Q_{n t_i, h})\|^2 + v(Q_{n t_i, h}).$$

This result is a direct consequence of Proposition 9. Therefore, Algorithm 1 provides an approximation of the k -PDTM. Since the algorithm does not converge to the optimal centers, we suggest running the algorithm several times and storing the best solution in terms of the empirical cost R_n , as for k -means.

As mentioned above, Algorithm 9 may be thought of as a special instance of Lloyd’s algorithm [32]. This algorithm consists in repeatedly decomposing the space \mathbb{R}^d into cells associated to the t_i ’s, and then replacing the t_i ’s by the means of P restricted to the cells. This kind of algorithm provably outputs local minimizers of risks of the form $R_d : \boldsymbol{\tau} \mapsto P \min_{i \in \llbracket 1, k \rrbracket} d(\cdot, \tau_i)$, for any Bregman divergence d (see, e.g., [6]). We recall that a Bregman divergence is defined by $d(x, y) = \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle$, for some convex function ϕ . Actually, Lloyd’s algorithm only works for Bregman divergences [5], since they are the only functionals d such that $c \mapsto P d(\cdot, c)$ attains its minimum at $c = P \cdot$, the expectation of P . This suggests that our criterion R may be expressed in terms of some Bregman divergence.

For $P \in \mathcal{P}_2(\mathbb{R}^d)$, according to [16, Proposition 3.6], the function $\psi_{P, h} : x \mapsto \|x\|^2 - d_{P, h}^2(x)$ is convex, and its set of subgradients at x is given by $\Delta_{x, h} = \{2m(P_{x, h}) \mid P_{x, h} \in \mathcal{P}_{x, h}(P)\}$. A simple computation based on (1) shows that the Bregman divergence associated with $\psi_{P, h}$ is defined for every $x, t \in \mathbb{R}^d$ by

$$d_{\psi_{P, h}}(x, t) = \|x - m(P_{t, h})\|^2 + v(P_{t, h}) - d_{P, h}^2(x). \quad (9)$$

Since $d_{P, h}^2(x)$ does not depend on t , our criterion R has the same minimizers as R_d for the Bregman divergence $d = d_{\psi_{P, h}}$. Thus, Proposition 9 is a consequence of the fact that $\psi_{P, h}$ is a Bregman divergence.

4 Numerical illustrations

4.1 Topological inference from noisy pointclouds

Let M be a compact subset of \mathbb{R}^d . Geometric and topological information about M can be recovered from some r -sublevel sets of the function distance to M , d_M (see, e.g., [16, Proposition 4.3]). To tackle the tough question of the selection of r , or simply to track multiscale information, the concept of persistent homology has been introduced in [23]. It consists in describing the evolution of the homology (number of connected components, holes, etc.) of the sublevel sets of d_M . Persistent homology can be encoded via persistence diagrams. A persistence diagram is a multiset of points (b, d) . Each point (b, d) is associated to one topological feature (a connected component, a hole, a void, etc.) that appears when $r = b$ (its birth time) and disappears when $r = d$ (its death

time). As well, if $\|\hat{d} - d_M\|_\infty$ is small enough, then the persistence diagrams associated with \hat{d} and d_M will be provably close [19], that is the lifetimes $d - b$ of the topological features will be similar. To assess the relevancy of our approach in a noisy topological inference setting, we will compute the persistence diagrams associated with the empirical k -PDTM, and compare it with the outputs of other methods.

Following [28], we choose for M the infinity symbol embedded in \mathbb{R}^2 . The persistence diagram associated to d_M is depicted in Figure 2 (right). This diagram contains one red point $(0, \infty)$, that corresponds to the connected component (0-dimensional topological feature), and two green points that correspond to the two holes (1-dimensional topological features). We generated a sample of 200 points, uniformly on the infinity symbol, with an additional additive Gaussian noise, with standard deviation $\sigma = 0.02$. This sample is corrupted by 80 outliers – 40 points generated according to the uniform distribution on the rectangle $[-2, 5] \times [-2, 2]$ and 40 points on the rectangle $[-4, 7] \times [-4, 4]$. This results in a corrupted sample \mathbb{X}_n of 280 points.

We compared three methods to recover relevant features of M from \mathbb{X}_n . Each method boils down to build an approximation f of d_M . These functions are of the type $f : x \mapsto \sqrt{\min_{i \in I} \|x - \tau_i\|^2 + \omega_i^2}$, for some finite set I , centers $\tau_i \in \mathbb{R}^d$ and weights $\omega_i \geq 0$. The first function we consider is derived from the k -means algorithm [34] ($|I| = k$, centers τ_i are given by the optima of the k -means criterion and $\omega_i = 0$), the second is the q -witnessed distance [28] ($|I| = n = 280$, it coincides with the k -PDTM for $k = n$, with mass parameter $h = q/n$) and the third one is the k -PDTM ($|I| = k$, with mass parameter $h = q/n$).

These methods depend on two parameters q and k . Providing a method to calibrate q and k in general is beyond the scope of the paper. Here, we choose $q = 10$ and $k = 50$. Roughly, q is chosen small enough so that the distance to the q -th nearest neighbor remains small compared to the curvature of M but large enough to deal with noise, and k is chosen large enough so that a uniform grid with k points has grid size small compared to the curvature of M . More details on this heuristic can be found in the Appendix.

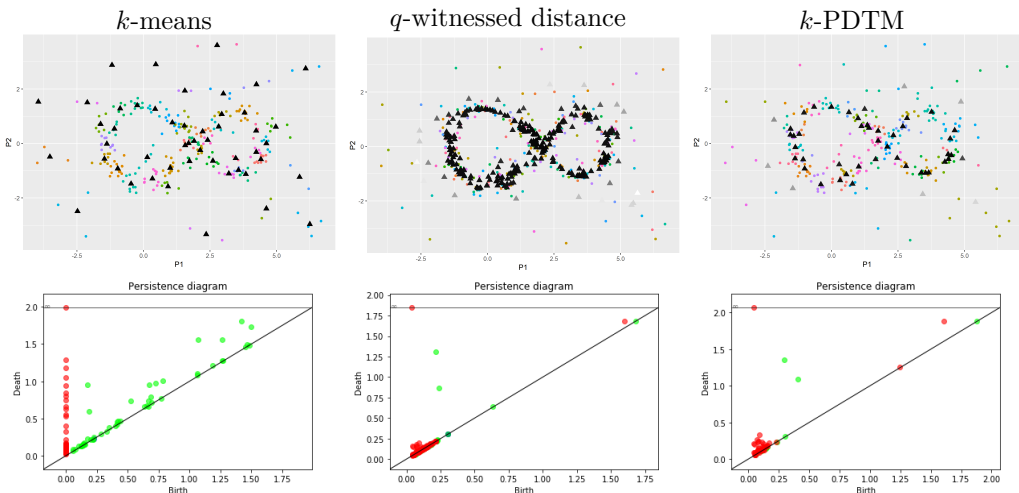


Figure 1: Comparison of the basic methods

We implemented the three methods with the R software. We used the R function `kmeans` for the first method, and the `FNN` R library to compute the nearest neighbors for the two other methods. In Figure 1 we plotted the points of \mathbb{X}_n . Points are represented with the same color when they lie in the same weighted Voronoi cell (for the centers τ_i and weights ω_i^2). Centers τ_i are represented by triangles and colored in function of the

weights ω_i^2 (black centers correspond to $\omega_i^2 = 0$). The second row of Figure 1 depicts the corresponding persistence diagrams, for the three methods. They were obtained using the function `weighted_alpha_complex_3d_persistence` in the `Gudhi C++` library, based on alpha-shapes [22]. We observe on Figure 1 that the three main features of the symbol infinity (one connected component, two holes) are recovered for the k -PDTM and the q -witnessed distance, but not for k -means. As exposed in Lemma 4, this is due to the "void-filling" drawback of k -means.

4.2 Outliers detection

Another possible interest of the proposed method is outlier detection, based on the following principle: if an observation X_i is such that $d_M(X_i)$ is large, then it can be considered as an outlier. Replacing d_M with an approximation gives the intuition of our denoising scheme.

Such a procedure need as an input a *level* α , that is the proportion of points that will be considered as signal points. Note that there exist heuristics to empirically design such an α (see, e.g., [13]). A level α being given, a straightforward approach consists in removing the $n(1 - \alpha)$ points that corresponds to the largest values of \hat{d} , for an estimate \hat{d} of d_M . In the following, we refer to this method as *truncation*, resulting in *truncated k -means*, *truncated q -witnessed* and *truncated k -PDTM*.

However, it is possible to combine compression and denoising, by looking simultaneously for a subset of $n\alpha$ points (*trimming set*) and a set of k points that approximates the best the trimming set. For a non-negative function d , this corresponds to the minimization of the criterion $\tau \mapsto \inf_{P_\alpha \in \mathcal{P}_\alpha(P)} P_\alpha \min_{i \in [1, k]} d(\cdot, \tau_i)$. Whenever d is a Bregman divergence, minimizers of such a criterion may be obtained via a Lloyd-type algorithm (see, e.g., [13]). Fortunately, since the k -means distance and the k -PDTM may be expressed via Bregman divergences, namely the squared Euclidean norm and (9), the procedure exposed in [13] applies. The outputs of the aforementioned procedure will be called *trimmed k -means* ([20]) and *trimmed k -PDTM*. We experiment each of these methods for the dataset of the previous section (200 signal points around the infinity symbol, 80 ambient noise points). We choose $\alpha = 200$, $q = 10$ and $k = 50$. Figure 2 depicts, for the trimmed versions of the algorithms, the resulting partition signal/outliers along with the k centers and weights (shade of triangles) in the first row. The second row exposes the corresponding persistence diagrams. Similar illustrations for the *truncated* algorithms may be found in the Appendix.

As depicted by Figure 2, the trimmed k -PDTM globally succeeds in identifying noise points and providing a relevant geometric approximation of the signal, whereas trimmed k -means still adds one spurious 1-dimensional topological feature. To be more precise on the topological performances of the aforementioned methods, we repeated the experiment 100 times. At each time, we computed the lifetimes of the topological features and sorted them in decreasing order. Figure 3 below exposes the means of these lifetimes. We see that k -meanbased methods add spurious holes (corresponding to the three last 1-dimensional features), and add many spurious connected components. Note that the first 0-dimensional feature, corresponding to the infinite connected component, has been removed in Figure 3. On the whole, our method compares well with q -witnessed-based methods.

The diagram on the right of Figure 3 depicts the mean amount of False positive over the 100 repetitions, that is the number of signal points that are labeled as outliers by the algorithm. We also include comparison with other trimming approaches for outlier detection, such as `tclust` [26] (`tclust` function in `trimcluster` R library) and the truncated version of k -median [15] (`kGmedian` function of the `Gmedian` R library). Again, our method compares well with q -witnessed-based denoising, contrary to the other

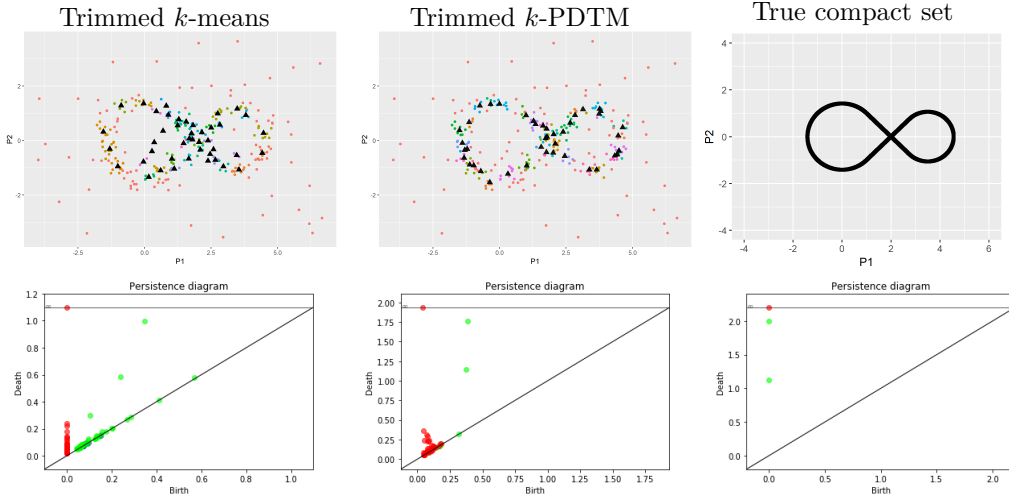


Figure 2: Comparison of the trimmed versions of the methods

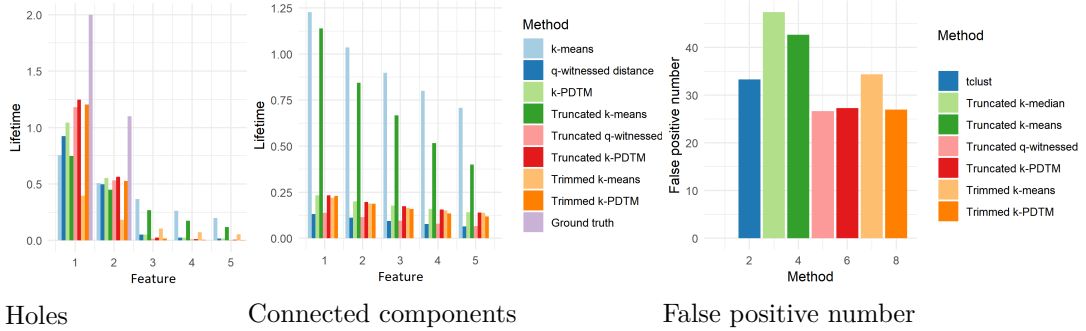


Figure 3: Features lifetimes and False positive number

Euclidean-based methods.

5 Proofs for Section 2

5.1 Proof of Proposition 6

It suffices to prove that $\mathcal{P}_h(P)$ is a compact set (for the weak convergence metric) and that R is continuous (i.e. that $P_h \mapsto m(P_h)$ and $P_h \mapsto M(P_h)$ are continuous on $\mathcal{P}_h(P)$).

The set $h\mathcal{P}_h(P)$ is tight. Prokhorov's Theorem entails that, for any sequence $(\mu_n/h)_{n \in \mathbb{N}}$ in $\mathcal{P}_h(P)$, up to a subsequence, there exists μ a Borel positive measure on \mathbb{R}^d such that μ_n converges weakly to μ . The dominated convergence Lemma applied to the functions $\mathbb{1}_{\mathbb{R}^d}$ and $\mathbb{1}_O$, for an open set O , ensures that $\mu(\mathbb{R}^d) = h$ and $\mu(O) \leq P(O)$, proving that $\mu \in h\mathcal{P}_h(P)$. Then, [4, p.438] yields that μ and P are regular measures. Thus, μ is a submeasure of P of mass h , and $\mathcal{P}_h(P)$ is compact.

We will now prove that the maps $P_h \mapsto m(P_h)$ and $P_h \mapsto M(P_h)$ are continuous on $\mathcal{P}_h(P)$. For $M > 0$ and $u \in \mathbb{R}^d$, denote by $u \wedge M$ the vector $(u_1 \wedge M, \dots, u_d \wedge M)$, where $u_i \wedge M$ denotes $\min(u_i, M)$. Consider $(P_{h,n})_{n \in \mathbb{N}}$ a sequence in $\mathcal{P}_h(P)$ converging to some distribution P_h . Then there exists $M_\varepsilon > 0$ such that for every $P'_h \in \mathcal{P}_h(P)$, $\|P'_h(\cdot \wedge M_\varepsilon) - P_h(\cdot \wedge M_\varepsilon)\| \leq P(\|\cdot\|_\infty > M_\varepsilon)/h \leq \varepsilon$. On the other hand, since $\cdot \mapsto \cdot \wedge M_\varepsilon$ is bounded and continuous, $\|P_{h,n}(\cdot \wedge M_\varepsilon) - P_h(\cdot \wedge M_\varepsilon)\|$ converges to 0. This proves the

continuity of $P_h \mapsto m(P_h)$. We also have that $P_{h,n} \|\cdot\|^2 \wedge M \rightarrow P_h \|\cdot\|^2 \wedge M$. Since for every $P'_h \in \mathcal{P}_h(P)$, $|P'_h(\|\cdot\|^2 \wedge M - \|\cdot\|^2)| \leq (P \|\cdot\|^2 \mathbb{1}_{\|\cdot\|^2 > M})/h$ and P has a finite second order moment, we deduce as well that $M(P_{h,n}) \rightarrow M(P_h)$. \square

5.2 Proof of Proposition 9

For short, we use the notation $m_i = m(\tilde{P}_{i,h})$, $v_i = v(\tilde{P}_{i,h})$ and $Q(du)f(u)$ for the expectation of f with respect to the Borel measure Q . Then, a bivariance decomposition yields

$$\begin{aligned} R(P_1, \dots, P_k) &= P(du) \min_{i \in [1, k]} P_i(dz) \|u - z\|^2 = \sum_{i=1}^k \tilde{P}_{i,h}(du) P_i(dz) \|u - z\|^2 \\ &= \sum_{i=1}^k \tilde{P}_{i,h}(\mathbb{R}^d) P_i(dz) (\|z - m_i\|^2 + v_i) \geq \sum_{i=1}^k \tilde{P}_{i,h}(\mathbb{R}^d) P_{m_i, h}(dz) (\|z - m_i\|^2 + v_i) \\ &= \sum_{i=1}^k \tilde{P}_{i,h}(du) P_{m_i, h}(dz) \|z - u\|^2, \end{aligned}$$

where $P_{m_i, h} \in \mathcal{P}_{m_i, h}(P)$, and equality holds if and only if $P_i \in \mathcal{P}_{m_i, h}(P)$. Thus, denoting by $(\tilde{P}_{m_i, h})_{i \in [1, k]}$ the set of weighted Voronoi measures associated to the measures $(P_{m_i, h})_{i \in [1, k]}$, we have

$$\begin{aligned} R(P_1, \dots, P_k) &\geq \sum_{i=1}^k \tilde{P}_{m_i, h}(du) P_{m_i, h}(dz) \|z - u\|^2 \\ &= \sum_{i=1}^k \tilde{P}_{m_i, h}(du) (\|m(P_{m_i, h}) - u\|^2 + v(P_{m_i, h})) = R(P_{m_1, h}, \dots, P_{m_k, h}). \end{aligned}$$

\square

5.3 Proof of Lemma 11

Let $g(x, P_h) = M(P_h) - \|\tau\|^2 + 2 \langle x, \tau - m(P_h) \rangle$. Then (4) entails that $\omega_{P, h}^2(\tau) = \sup_{x \in \mathbb{R}^d} \inf_{P_h \in \mathcal{P}_h(P)} g(x, P_h)$. According to Section 5.1, $\mathcal{P}_h(P)$ is a compact set, and $P_h \mapsto m(P_h)$ as well as $P_h \mapsto M(P_h)$ are continuous. So, for every $x \in \mathbb{R}^d$, $g(x, \cdot)$ is continuous, and linear. On the other hand, for every P_h in $\mathcal{P}_h(P)$, $g(\cdot, P_h)$ is linear and continuous. Sion's theorem [30] yields that

$$\omega_{P, h}^2(\tau) = \min_{P_h \in \mathcal{P}_h(P)} \sup_{x \in \mathbb{R}^d} M(P_h) - \|\tau\|^2 + 2 \langle x, \tau - m(P_h) \rangle. \quad (10)$$

Therefore, $\omega_{P, h}^2(\tau) < \infty$ is equivalent to $\tau \in \tilde{\mathcal{M}}_h(P)$. Now let τ be in $\tilde{\mathcal{M}}_h(P)$. According to (10), we have

$$\omega_{P, h}^2(\tau) = \inf_{P_h \in \mathcal{P}_h(P), m(P_h) = \tau} M(P_h) - \|\tau\|^2 = \inf_{P_h \in \mathcal{P}_h(P), m(P_h) = \tau} v(P_h).$$

Since $P_h \mapsto v(P_h)$ is continuous on $\mathcal{P}_h(P)$ and $\mathcal{P}_h(P) \cap m^{-1}(\{\tau\})$ is compact, there exists P_h such that $m(P_h) = \tau$ and $\omega_{P, h}^2(\tau) = v(P_h)$. \square

5.4 Proof of Theorem 12

Let \tilde{R} denote $\boldsymbol{\tau} \mapsto P \min_{j \in [1, k]} \|\cdot - \tau_j\|^2 + \omega_{P, h}^2(\tau_j)$, for $\boldsymbol{\tau} \in (\mathbb{R}^d)^{(k)}$. According to Lemma 11, $\boldsymbol{\tau}^*$ is a minimizer of \tilde{R} if and only if $\tau_i^* = m(P_i^*)$ and $\omega_{P, h}^2(\tau_i^*) = v(P_i^*)$ for (P_1^*, \dots, P_k^*) minimizing R . According to Proposition 9, for such a $\boldsymbol{\tau}^*$, with the notation $m_i = m(\tilde{P}_{i, h}^*)$, we have $P_i^* \in \mathcal{P}_{m_i, h}(P)$. Hence, $\tau_i^* = m(P_{m_i, h})$ and $\omega_{P, h}^2(\tau_i^*) = v(P_{m_i, h})$, for some $P_{m_i, h} \in \mathcal{P}_{m_i, h}(P)$, and $\mathbf{t}^* = (m_1, \dots, m_k)$ is an R -minimizer. Conversely, Proposition 9 ensures that if \mathbf{t}^* is an R -minimizer, then $\boldsymbol{\tau}^* = (\tau_i^*)_{i \in [1, k]}$, with $\tau_i^* = m(P_{t_i^*, h})$, is a \tilde{R} -minimizer. \square

6 Proofs for Section 3

6.1 Intermediate results

The proofs of the Section 3 results will make intensive use of the following Lemmas, whose proofs are postponed to the Appendix. We first mention some known results about sub-Gaussian distributions.

Lemma 23.

Let $Q \in \mathcal{P}^{(V)}(\mathbb{R}^d)$, a sub-Gaussian measure with variance $V^2 > 0$, and $Q_h \in \mathcal{P}_h(Q)$. Then we have

$$Q_h \|\cdot\|^2 \leq \frac{3V^2}{h}.$$

The proof of Lemma 23 is given in Section A.2 of the Appendix. Next, Lemma 24 below ensures that the distributions involved in Theorem 19 are sub-Gaussian.

Lemma 24.

If Y is a random variable sampled from a distribution P in $\mathcal{P}^K(\mathbb{R}^d)$ and Z is independent from Y and sampled from a distribution Q' in $\mathcal{P}^{(\sigma)}(\mathbb{R}^d)$ for some $\sigma > 0$. Then, the distribution Q of the random variable $X = Y + Z$ is sub-Gaussian with variance $V^2 = (K + \sigma)^2$, that is in $\mathcal{P}^{(K+\sigma)}(\mathbb{R}^d)$.

Moreover,

$$W_1(P, Q) \leq 3\sigma \text{ and } W_2(P, Q) \leq \sqrt{3}\sigma.$$

A proof of Lemma 24 can be found in Section A.3, Appendix. In what follows, we let γ and $\hat{\gamma}$ denote the functions

$$\begin{aligned} \gamma(\mathbf{t}, x) &= \min_{i \in [1, k]} -2\langle x, m(Q_{t_i, h}) \rangle + \|m(Q_{t_i, h})\|^2 + v(Q_{t_i, h}), \\ \hat{\gamma}(\mathbf{t}, x) &= \min_{i \in [1, k]} -2\langle x, m(Q_{n t_i, h}) \rangle + \|m(Q_{n t_i, h})\|^2 + v(Q_{n t_i, h}), \end{aligned} \quad (11)$$

for $(\mathbf{t}, x) \in (\mathbb{R}^d)^{(k)} \times \mathbb{R}^d$ with $\mathbf{t} = (t_1, t_2, \dots, t_k)$. We will use two deviation bounds, stated below.

Lemma 25.

If Q is sub-Gaussian with variance V^2 , then, for every $p > 0$, with probability larger than $1 - n^{-p}$, we have

$$\sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} |(Q - Q_n)\gamma(\mathbf{t}, \cdot)| \leq C \frac{V^2 \sqrt{kd}(1+p)^{\frac{3}{2}} \log(n)^{\frac{3}{2}}}{h\sqrt{n}},$$

for some absolute positive constant C .

The proof of Lemma 25 is deferred to Section B.5 of the Appendix.

Lemma 26.

Assume that Q is sub-Gaussian with variance V^2 , then, for every $p > 0$, with probability larger than $1 - 9n^{-p}$, we have

$$\begin{aligned} \sup_{t \in \mathbb{R}^d} \|m(Q_{t,h}) - m(Q_{n t,h})\| &\leq \frac{CV\sqrt{d}(p+1)\log(n)}{h\sqrt{n}}, \\ \sup_{t \in (\mathbb{R}^d)^{(k)}} |Q_n(\gamma - \hat{\gamma})(t, \cdot)| &\leq CV^2 \frac{\sqrt{d}(p+1)^{\frac{3}{2}}\log(n)^{\frac{3}{2}}}{h\sqrt{n}}. \end{aligned}$$

As well, the proof of Lemma 26 is deferred to Section B.4 in the Appendix.

6.2 Proof of Proposition 14

The first inequality comes from (4). We now focus on the second bound. By definition of $d_{P,h,k}$, for all $\mathbf{t} = (t_1, t_2, \dots, t_k) \in (\mathbb{R}^d)^{(k)}$ we have $Pd_{P,h,k}^2 \leq P \min_{i \in [1,k]} \|\cdot - m(P_{t_i,h})\|^2 + v(P_{t_i,h})$. Thus,

$$\begin{aligned} P(d_{P,h,k}^2 - d_{P,h}^2) &\leq P \left(\min_{i \in [1,k]} \|\cdot - m(P_{t_i,h})\|^2 + v(P_{t_i,h}) - d_{P,h}^2 \right) \\ &= P \left(\min_{i \in [1,k]} (d_{P,h}^2(t_i) - \|t_i\|^2) - (d_{P,h}^2 - \|\cdot\|^2) + \langle \cdot - t_i, -2m(P_{t_i,h}) \rangle \right), \end{aligned}$$

according to (1). Now [16, Corollary 3.7] ensures that, for x, y in \mathbb{R}^d ,

$$\|y\|^2 - d_{P,h}^2(y) - (\|x\|^2 - d_{P,h}^2(x)) \geq 2 \langle m(P_{x,h}), y - x \rangle. \quad (12)$$

We deduce that

$$\begin{aligned} P(d_{P,h,k}^2 - d_{P,h}^2) &\leq P \min_{i \in [1,k]} 2 \langle \cdot - t_i, m(P_{\cdot,h}) - m(P_{t_i,h}) \rangle \\ &\leq 2P \min_{i \in [1,k]} \|\cdot - t_i\| \|m(P_{\cdot,h}) - m(P_{t_i,h})\|. \end{aligned}$$

Now choose t_1, \dots, t_k such that $M \subset \bigcup_{i \in [1,k]} B(t_i, f_M^{-1}(k))$. The result follows. \square

6.3 Proof of Corollary 15

The proof of Corollary 15 is based on the following bounds, in the case where P is absolutely continuous with respect to the Lebesgue measure, with density f satisfying $0 < f_{\min} \leq f \leq f_{\max}$.

$$f_M^{-1}(k) \leq 2K\sqrt{d}k^{-1/d}, \quad (13)$$

$$\zeta_{P,h}(f_M^{-1}(k)) \leq C_{f_{\max}, K, d, h} k^{-1/d}. \quad (14)$$

First, note that since $M \subset B(0, K)$, for any $\varepsilon > 0$, $f_M(\varepsilon) \leq f_{B(0,K)}(\varepsilon) \leq \left(\frac{2K\sqrt{d}}{\varepsilon}\right)^d$, hence (13). To prove the second inequality, we have to give a bound on the modulus of continuity $\zeta_{P,h}$. Let x, y be in M , and denote by $\delta = \|x - y\|$. Since P has a density, $P\partial B(x, \delta_{P,h}(x)) = P\partial B(y, \delta_{P,h}(y)) = 0$. We deduce that $P_{x,h} = \frac{1}{h}P|_{B(x, \delta_{P,h}(x))}$ and $P_{y,h} = \frac{1}{h}P|_{B(y, \delta_{P,h}(y))}$. Without loss of generality, assume that $\delta_{P,h}(x) \geq \delta_{P,h}(y)$. Then

$B(y, \delta_{P,h}(y)) \subset B(x, \delta_{P,h}(x) + \delta)$. We may bound $\|m(P_{x,h}) - m(P_{y,h})\|$ by

$$\begin{aligned} & \frac{1}{h} \left\| P \cdot (\mathbb{1}_{B(x, \delta_{P,h}(x))} - \mathbb{1}_{B(y, \delta_{P,h}(y))}) \right\| \leq \frac{K}{h} P \left| \mathbb{1}_{B(x, \delta_{P,h}(x))} - \mathbb{1}_{B(y, \delta_{P,h}(y))} \right| \\ & = 2 \frac{K}{h} P (B(y, \delta_{P,h}(y)) \setminus (B(x, \delta_{P,h}(x)) \cap B(y, \delta_{P,h}(y)))) \\ & \leq 2 \frac{K}{h} P (B(x, \delta_{P,h}(x) + \delta) \cap B(x, \delta_{P,h}(x))^c) \\ & = 2 \frac{K}{h} \omega_d [(\delta_{P,h}(x) + \delta)^d - \delta_{P,h}(x)^d] \leq 2 \frac{K^{d+1} \omega_d}{h} \left[\left(1 + \frac{\delta}{\delta_{P,h}(x)} \right)^d - 1 \right], \end{aligned}$$

where ω_d denotes the Lebesgue volume of the ball $B(0, 1)$ in \mathbb{R}^d . Since $(1+v)^d \leq 1 + d(1+v)^{d-1}v$, for $v \geq 0$, and $\delta_{P,h}(x) \geq \left(\frac{h}{f_{max} \omega_d} \right)^{1/d}$, we have $\zeta_{P,h}(\delta) \leq C_{f_{max}, K, d, h} \delta$, hence (14). The result of Corollary 15 follows. \square

6.4 Proof of Corollary 16

Since N is a \mathcal{C}^2 -submanifold, its reach ρ (as defined in [25, Definition 4.1]) is positive. Without loss of generality we assume that N is connected. Since P has a density with respect to the volume measure on N , we have $P(N^\circ) = 1$. Thus we take $M = N^\circ$, that is the set of interior points. Since P satisfies a (cf_{min}, d') -standard assumption, we have $f_M(\varepsilon) \leq 2^{d'} / (cf_{min} r^{-d'})$, according to [18, Lemma 10]. Hence $f_M^{-1}(k) \leq C_{f_{min}, N} k^{-1/d'}$.

It remains to bound the continuity modulus of $x \mapsto m(P_{x,h})$. For any x in M , since $P(\partial N) = 0$ and P has a density with respect to the volume measure on N , we have $P_{x,h} = P|_{B(x, \delta_{P,h}(x))}$. Besides, since for all $r > 0$, $P(B(x, r)) \geq cf_{min} r^{d'}$, we may write $\delta_{P,h}(x) \leq c_{N, f_{min}} h^{1/d'} \leq \rho/12$, for h small enough. Now let x and y be in M so that $\|x - y\| = \delta \leq \rho/12$, and without loss of generality assume that $\delta_{P,h}(x) \geq \delta_{P,h}(y)$. Then, proceeding as in the proof of (14), it comes

$$\|m(P_{x,h}) - m(P_{y,h})\| \leq \frac{2K}{h} P (B(x, \delta_{P,h}(x) + \delta) \cap B(x, \delta_{P,h}(x))^c).$$

Since $\delta_{P,h}(x) + \delta \leq \rho/6$, for any u in $B(x, \delta_{P,h}(x) + \delta) \cap M$ we may write $u = \exp_x(rv)$, where $v \in T_x M$ with $\|v\| = 1$ and $r = d_N(u, x)$ is the geodesic distance between u and x (see, e.g., [25, Theorem 4.18] or [2, Proposition 25]). Note that, according to [2, Proposition 26], for any u_1 and u_2 such that $\|u_1 - u_2\| \leq \rho/4$,

$$\|u_1 - u_2\| \leq d_N(u_1, u_2) \leq 2\|u_1 - u_2\|. \quad (15)$$

Now let p_1, \dots, p_m be a δ -covering of the sphere $S(x, \delta_{P,h}(x)) = \overline{B}(x, \delta_{P,h}(x)) \setminus B(x, \delta_{P,h}(x))$. According to (15), we may choose $m \leq c_{d'} \delta_{P,h}(x)^{d'-1} \delta^{-(d'-1)}$.

Now, for any u such that $u \in M$ and $\delta_{P,h}(x) \leq \|x - u\| \leq \delta_{P,h}(x) + \delta$, there exists $t \in S(x, \delta_{P,h}(x))$ such that $\|t - u\| \leq 2\delta$. Hence

$$P (B(x, \delta_{P,h}(x) + \delta) \cap B(x, \delta_{P,h}(x))^c) \leq \sum_{j=1}^m P (B(p_j, 2\delta)).$$

For any j , since $2\delta \leq \rho/6$, in local polar coordinates around p_j we may write,

$$P (B(p_j, 2\delta)) \leq \int_{\{r, v \mid \exp_{p_j}(rv) \in M, r \leq 4\delta\}} f(r, v) J(r, v) dr dv \leq f_{max} \int_{\{r, v \mid r \leq 4\delta\}} J(r, v) dr dv,$$

using (15), where $J(r, v)$ denotes the Jacobian of the volume form. According to [2, Proposition 27], we have $J(r, v) \leq C_{d'} r^{d'}$, hence $P(B(p_j, 2\delta)) \leq C_{d'} f_{max} \delta^{d'}$. Thus,

$$\|m(P_{x,h}) - m(P_{y,h})\| \leq \frac{2K}{h} m C_{d'} f_{max} \delta^{d'} \leq C_{N, f_{max}, f_{min}} \delta.$$

Choosing k large enough so that $f_M^{-1}(k) \leq C_{f_{min}, N} k^{-1/d'} \leq \rho/12$ gives the result. \square

6.5 Proof of Proposition 17

For all $x \in \text{Supp}(P)$,

$$\begin{aligned} d_{Q,h,k}^2(x) - d_{P,h}^2(x) &= d_{Q,h,k}^2(x) - d_{Q,h}^2(x) + d_{Q,h}^2(x) - d_{P,h}^2(x) \\ &\geq -\|d_{P,h}^2 - d_{Q,h}^2\|_{\infty, \text{Supp}(P)}. \end{aligned}$$

Thus, $(d_{Q,h,k}^2 - d_{P,h}^2)_- \leq \|d_{P,h}^2 - d_{Q,h}^2\|_{\infty, \text{Supp}(P)}$ on $\text{Supp}(P)$, where $f_- : x \mapsto f(x) \mathbb{1}_{f(x) \leq 0}$ denotes the negative part of any function f on \mathbb{R}^d . Then,

$$\begin{aligned} P |d_{Q,h,k}^2 - d_{P,h}^2| &= P(d_{Q,h,k}^2 - d_{P,h}^2) + 2(d_{Q,h,k}^2 - d_{P,h}^2)_- \\ &\leq P\Delta + P(d_{P,h}^2 - d_{Q,h}^2) + 2\|d_{P,h}^2 - d_{Q,h}^2\|_{\infty, \text{Supp}(P)}, \end{aligned}$$

with $\Delta = d_{Q,h,k}^2 - d_{P,h,k}^2$. To bound $P\Delta$ from above, let $\mathbf{s} \in (\overline{B}(0, K))^{(k)}$ be a k -points minimizer of R for P , such that when $\tilde{P}_{s_i, h}(\mathbb{R}^d) \neq 0$, $s_i = m(\tilde{P}_{s_i, h})$. Such an \mathbf{s} exists according to Proposition 9 and Lemma 11. Set $f_{Q,t}(x) = -2\langle x, m(Q_{t,h}) \rangle + M(Q_{t,h})$ for $t \in \mathbb{R}^d$, and let \mathbf{t} be a k -points minimizer of R for Q .

$$\begin{aligned} P\Delta &= P \left(\min_{i \in [1, k]} f_{Q, t_i} - \min_{i \in [1, k]} f_{P, s_i} \right) \\ &\leq (P - Q) \min_{i \in [1, k]} f_{Q, t_i} + (Q - P) \min_{i \in [1, k]} f_{Q, s_i} + P \left(\min_{i \in [1, k]} f_{Q, s_i} - \min_{i \in [1, k]} f_{P, s_i} \right). \end{aligned}$$

For a transport plan π between P and Q , $(P - Q) \min_{i \in [1, k]} f_{Q, t_i}$ is bounded by

$$\begin{aligned} \mathbb{E}_{(X, Y) \sim \pi} \left[\min_{i \in [1, k]} -2\langle X, m(Q_{t_i, h}) \rangle + M(Q_{t_i, h}) - \min_{i \in [1, k]} -2\langle Y, m(Q_{t_i, h}) \rangle + M(Q_{t_i, h}) \right] \\ \leq 2\mathbb{E}_{(X, Y) \sim \pi} \left[\sup_{t \in \mathbb{R}^d} \langle Y - X, m(Q_{t, h}) \rangle \right]. \end{aligned}$$

Thus, $(P - Q) \min_{i \in [1, k]} f_{Q, t_i} \leq 2W_1(P, Q) \sup_{t \in \mathbb{R}^d} \|m(Q_{t, h})\|$, choosing for π the optimal transport plan for the W_1 distance between P and Q in (3). Also note that $P(\min_{i \in [1, k]} f_{Q, s_i} - \min_{i \in [1, k]} f_{P, s_i})$ is bounded from above by

$$\begin{aligned} &\sum_{i=1}^k \tilde{P}_{s_i, h} (-2\langle \cdot, m(Q_{s_i, h}) \rangle + M(Q_{s_i, h})) - (-2\langle \cdot, m(P_{s_i, h}) \rangle + M(P_{s_i, h})) \\ &= \sum_{i=1}^k \tilde{P}_{s_i, h} 2\langle -s_i, m(P_{s_i, h}) - m(Q_{s_i, h}) \rangle + d_{Q,h}^2(s_i) - d_{P,h}^2(s_i) \\ &\leq \|d_{P,h}^2 - d_{Q,h}^2\|_{\infty, B(0, K)} + 2 \sum_{i=1}^k \tilde{P}_{s_i, h}(\mathbb{R}^d) \langle m(\tilde{P}_{s_i, h}) - s_i, m(P_{s_i, h}) - m(Q_{s_i, h}) \rangle. \end{aligned}$$

Since $s_i = m(\tilde{P}_{s_i, h})$, the result follows. \square

6.6 Proof of Proposition 18

Let $\Delta_{\infty,K}$ denote $\sup_{x \in M} d_{Q,h,k}(x)$, and let $x \in M$ achieving the maximum distance. Since $d_{Q,h,k}$ is 1-Lipschitz, we deduce that $B\left(x, \frac{\Delta_{\infty,K}}{2}\right) \subset \left\{y \mid d_{Q,h,k}(y) \geq \frac{\Delta_{\infty,K}}{2}\right\}$. Since $P\left(B\left(x, \frac{\Delta_{\infty,K}}{2}\right)\right) \geq C(P) \left(\frac{\Delta_{\infty,K}}{2}\right)^{d'} \wedge 1$, Markov inequality yields that

$$\Delta_P^2 \geq C(P) \left(\frac{\Delta_{\infty,K}}{2}\right)^{d'+2} \wedge \frac{\Delta_{\infty,K}^2}{4}.$$

Thus we have $\sup_{x \in M} (d_{Q,h,k} - d_M)(x) = \Delta_{\infty,K} \leq C(P)^{-\frac{1}{d'+2}} \Delta_P^{\frac{2}{d'+2}} \vee 2\Delta_P$. Now, for $x \in \mathbb{R}^d$, we let $p \in M$ such that $\|x - p\| = d_M(x)$. Denote by $r = \|x - p\|$, and let t_j be such that $d_{Q,h,k}(p) = \sqrt{\|p - m(Q_{t_j,h})\|^2 + v(Q_{t_j,h})}$. Then

$$\begin{aligned} d_{Q,h,k}(x) &\leq \sqrt{\|x - m(Q_{t_j,h})\|^2 + v(Q_{t_j,h})} \\ &\leq \sqrt{d_{Q,h,k}^2(p) + r^2 + 2r\|p - m(Q_{t_j,h})\|} \\ &\leq \sqrt{d_{Q,h,k}^2(p) + r^2 + 2rd_{Q,h,k}(p)} \\ &= d_M(x) + (d_{Q,h,k}(p) - d_M(p)). \end{aligned}$$

Hence, $\sup_{x \in \mathbb{R}^d} (d_{Q,h,k} - d_M)(x) = \sup_{x \in M} (d_{Q,h,k} - d_M)(x) = \Delta_{\infty,K}$. On the other hand, we have $d_{Q,h,k} \geq d_{Q,h}$, along with $\|d_{Q,h} - d_{P,h}\|_{\infty} \leq h^{-\frac{1}{2}} W_2(P, Q)$ (see, e.g., [16, Theorem 3.5]) as well as $d_{P,h} \geq d_M$. Hence $d_{Q,h,k} \geq d_M - h^{-\frac{1}{2}} W_2(P, Q)$. \square

6.7 Proof of Theorem 19

We recall that γ and $\hat{\gamma}$ are defined in (11). According to Lemma 24, $Q \in \mathcal{P}^{(V)}(\mathbb{R}^d)$ with $V = \sigma + K$. Let

$$\begin{aligned} \mathbf{s} &= \arg \min \left\{ Q\gamma(\mathbf{t}, \cdot) \mid \mathbf{t} = (t_1, t_2, \dots, t_k) \in (\mathbb{R}^d)^{(k)} \right\}, \\ \hat{\mathbf{s}} &= \arg \min \left\{ Q_n \hat{\gamma}(\mathbf{t}, \cdot) \mid \mathbf{t} = (t_1, t_2, \dots, t_k) \in (\mathbb{R}^d)^{(k)} \right\}, \\ \tilde{\mathbf{s}} &= \arg \min \left\{ Q_n \gamma(\mathbf{t}, \cdot) \mid \mathbf{t} = (t_1, t_2, \dots, t_k) \in (\mathbb{R}^d)^{(k)} \right\}. \end{aligned}$$

With these notations, for all $x \in \mathbb{R}^d$, $d_{Q,h,k}^2(x) = \|x\|^2 + \gamma(\mathbf{s}, x)$ and $d_{Q_n,h,k}^2(x) = \|x\|^2 + \hat{\gamma}(\hat{\mathbf{s}}, x)$. We intend to bound $l(\mathbf{s}, \hat{\mathbf{s}}) = Q(d_{Q_n,h,k}^2 - d_{Q,h,k}^2) = Q(\gamma(\hat{\mathbf{s}}, \cdot) - Q\gamma(\mathbf{s}, \cdot))$.

$$\begin{aligned} l(\mathbf{s}, \hat{\mathbf{s}}) &= Q\gamma(\hat{\mathbf{s}}, \cdot) - Q_n\gamma(\hat{\mathbf{s}}, \cdot) + Q_n\gamma(\hat{\mathbf{s}}, \cdot) - Q_n\gamma(\tilde{\mathbf{s}}, \cdot) + Q_n\gamma(\tilde{\mathbf{s}}, \cdot) - Q\gamma(\mathbf{s}, \cdot) \\ &\leq \sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} (Q - Q_n)\gamma(\mathbf{t}, \cdot) + Q_n(\gamma - \hat{\gamma})(\hat{\mathbf{s}}, \cdot) \\ &\quad + Q_n(\hat{\gamma}(\hat{\mathbf{s}}, \cdot) - \hat{\gamma}(\tilde{\mathbf{s}}, \cdot)) + Q_n(\hat{\gamma} - \gamma)(\tilde{\mathbf{s}}, \cdot) + \sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} (Q_n - Q)\gamma(\mathbf{t}, \cdot), \end{aligned}$$

where we used $Q_n\gamma(\tilde{\mathbf{s}}, \cdot) \leq Q_n\gamma(\mathbf{s}, \cdot)$. Now, since $Q_n(\hat{\gamma}(\hat{\mathbf{s}}, \cdot) - \hat{\gamma}(\tilde{\mathbf{s}}, \cdot)) \leq 0$, we get

$$\begin{aligned} l(\mathbf{s}, \hat{\mathbf{s}}) &\leq \sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} (Q - Q_n)\gamma(\mathbf{t}, \cdot) + \sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} (Q_n - Q)\gamma(\mathbf{t}, \cdot) \\ &\quad + \sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} Q_n(\gamma - \hat{\gamma})(\mathbf{t}, \cdot) + \sup_{\mathbf{t} \in \mathbb{R}^d} Q_n(\hat{\gamma} - \gamma)(\mathbf{t}, \cdot). \end{aligned}$$

Combining Lemma 25 and Lemma 26 entails, with probability larger than $1 - 10n^{-p}$,

$$l(\mathbf{s}, \hat{\mathbf{s}}) \leq CV^2 \sqrt{kd} \frac{(p+1)^{\frac{3}{2}} \log(n)^{\frac{3}{2}}}{h\sqrt{n}}.$$

It remains to bound $|Pd_{Q_n, h, k}^2 - Qd_{Q_n, h, k}^2|$ as well as $|Pd_{Q, h, k}^2 - Qd_{Q, h, k}^2|$. To this aim we recall that $X = Y + Z$, Z being sub-Gaussian with variance σ^2 . Thus, denoting by $s_j(x) = \arg \min_{j \in [1, k]} \|x - m(Q_{s_j, h})\|^2 + v(Q_{s_j, h})$,

$$\begin{aligned} Pd_{Q, h, k}^2 - Qd_{Q, h, k}^2 &\leq \mathbb{E}_{(Y, Z)} [\|Y - m(Q_{s_j(Y+Z), h})\|^2 + v(Q_{s_j(Y+Z), h}) \\ &\quad - (\|Y + Z - m(Q_{s_j(Y+Z), h})\|^2 + v(Q_{s_j(Y+Z), h}))] \\ &\leq \mathbb{E}_Z \|Z\|^2 + 2\mathbb{E}_{(Y, Z)} \max_{j \in [1, k]} |\langle Z, m(Q_{s_j, h}) - Y \rangle| \\ &\leq 3\sigma^2 + 2\sqrt{3}\sigma \left(\max_{j \in [1, k]} \|m(Q_{s_j, h})\| + K \right) \leq \frac{C\sigma K}{\sqrt{h}}, \end{aligned}$$

using Cauchy-Schwarz inequality, Lemma 23 and $\sigma \leq K$. The converse bound on $Qd_{Q, h, k}^2 - Pd_{Q, h, k}^2$ may be proved the same way. Similarly, we may write

$$\begin{aligned} Pd_{Q_n, h, k}^2 - Qd_{Q_n, h, k}^2 &\leq 3\sigma^2 + 2\sqrt{3}\sigma \left(\max_{j \in [1, k]} \|m(Q_{n, s_j, h})\| + K \right) \\ &\leq 3\sigma^2 + 2\sqrt{3}\sigma \left(\max_{j \in [1, k]} \|m(Q_{s_j, h})\| + \sup_{t \in \mathbb{R}^d} \|m(Q_{t, h}) - m(Q_{n, t, h})\| + K \right) \\ &\leq 3\sigma^2 + 2\sqrt{3}\sigma \left(\max_{j \in [1, k]} \|m(Q_{s_j, h})\| + C(K + \sigma)\sqrt{d} \frac{(p+1)\log(n)}{h\sqrt{n}} + K \right) \\ &\leq \frac{C\sigma K}{\sqrt{h}} + \frac{C\sigma K\sqrt{d}(p+1)\log(n)}{h\sqrt{n}}, \end{aligned}$$

according to Lemma 23 and Lemma 26. The bound on $Qd_{Q_n, h, k}^2 - Pd_{Q_n, h, k}^2$ derives from the same argument. Collecting all pieces, we get, using $\sigma \leq K$,

$$\begin{aligned} |P(d_{Q_n, h, k}^2 - d_{Q, h, k}^2)| &\leq |Q(d_{Q_n, h, k}^2 - d_{Q, h, k}^2)| + \frac{C\sigma K\sqrt{d}(p+1)\log(n)}{h\sqrt{n}} + \frac{C\sigma K}{\sqrt{h}} \\ &\leq \frac{C\sigma K\sqrt{d}(p+1)\log(n)}{h\sqrt{n}} + \frac{CkK^2\sqrt{dk}((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + \frac{C\sigma K}{\sqrt{h}}. \end{aligned}$$

□

6.8 Proof of Proposition 20

Combining bounds obtained in Theorem 19 and Proposition 17 yields

$$\begin{aligned} |Pd_{Q_n, h, k}^2 - d_{P, h}^2| &\leq C\sqrt{kd} \frac{K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C \frac{K\sigma}{\sqrt{h}} \\ &\quad + 3\|d_{Q, h}^2 - d_{P, h}^2\|_{\infty, B(0, K)} + Pd_{P, h, k}^2 - d_{P, h}^2 + 4W_1(P, Q) \sup_{s \in \mathbb{R}^d} \|m(P_{s, h})\|. \end{aligned}$$

Using Corollary 16 and Lemma 23 entails

$$|Pd_{Q_n, h, k}^2 - d_{P, h}^2| \leq C\sqrt{kd} \frac{K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C \frac{K\sigma}{\sqrt{h}} + 3\|d_{Q, h}^2 - d_{P, h}^2\|_{\infty, B(0, K)} + C_P k^{-\frac{2}{d}}.$$

At last, using (2), Lemmas 23 and 24 leads to

$$\begin{aligned}
\|d_{Q,h}^2 - d_{P,h}^2\|_{\infty, B(0,K)} &\leq \|d_{Q,h} - d_{P,h}\|_{\infty, B(0,K)} (\|d_{Q,h}\|_{\infty, B(0,K)} + \|d_{P,h}\|_{\infty, B(0,K)}) \\
&\leq \frac{W_2(P, Q)}{\sqrt{h}} \left(\sup_{x \in B(0,K)} \sqrt{\|x - m(Q_{x,h})\|^2 + v(Q_{x,h})} + 2K \right) \\
&\leq \frac{\sqrt{3}\sigma}{\sqrt{h}} \left(\sqrt{K^2 + 2K\sqrt{3}\frac{\sigma + K}{\sqrt{h}} + 3\frac{(\sigma + K)^2}{h}} + 2K \right) \\
&\leq \frac{\sqrt{3}\sigma (3K + \sqrt{3}(K + \sigma))}{h}.
\end{aligned}$$

□

6.9 Proof of Proposition 21

As in the proof of [7, Theorem 1], for the sake of simplicity we assume that k is divisible by 3, set $m = (2k)/3$, and let z_1, \dots, z_m be a 6Δ -net in $B(0, K)$, with $\Delta = K/(6m^{\frac{1}{2}})$, so that such a net exists. We let as well w_1, \dots, w_m be in \mathbb{R}^d such that $\|w_i\| = \Delta$ and $z_i + w_i \in B(0, K)$. For $\sigma \in \{-1, +1\}^m$ such that $\sum_{i=1}^m \sigma_i = 0$ we denote by P_σ the distribution that satisfies, for $i \in \llbracket 1, m \rrbracket$,

$$P_\sigma(\{z_i\}) = P_\sigma(\{z_i + w_i\}) = \frac{(1 + \sigma_i \delta)}{2m},$$

with $\delta \leq \frac{1}{3}$. For $\tau \in \{-1, 1\}^{\frac{m}{2}}$, $\sigma(\tau)$ is defined by $\sigma(\tau)_j = \tau_j$ and $\sigma(\tau)_{\frac{m}{2}+j} = -\tau_j$, for $j \in \llbracket 1, m/2 \rrbracket$. We define now a p -points quantizer F as a map from \mathbb{R}^d such that $|F(\mathbb{R}^d)| = p$, and define F_σ as the k -points quantizer satisfying

$$\begin{aligned}
F_\sigma(z_i) &= z_i, & F_\sigma(z_i + w_i) &= z_i + w_i & \text{if } & \sigma_i = +1 \\
F_\sigma(z_i) &= F_\sigma(z_i + w_i) = z_i & & & \text{if } & \sigma_i = -1.
\end{aligned}$$

At last, for a quantizer F with images q_1, \dots, q_p and sets of preimages V_1, \dots, V_p , we denote by $R(F, P_\sigma)$ the quantity

$$R(F, P_\sigma) = \sum_{i=1}^p P_\sigma \left[\|\cdot - m(P_{q_i, h})\|^2 + v(P_{q_i, h}) \right] \mathbb{1}_{V_i},$$

where for $i \in \llbracket 1, p \rrbracket$, $P_{q_i, h} \in \mathcal{P}_{q_i, h}(P_\sigma)$. With a slight abuse we call nearest-neighbor quantizer a quantizer whose sets of preimages are the set of Voronoi cells associated with $(m(P_{q_i, h}), v(P_{q_i, h}))$, with ties arbitrarily broken.

The proof of Proposition 21 follows from the same arguments as [31, Proposition 3.1]. We first use the following Lemma.

Lemma 27.

Assume that $\delta \leq \frac{1}{3}$ and $h \leq \frac{1}{3m}$. Let σ and σ' be such that $\sum_{i=1}^m \sigma_i = \sum_{i=1}^m \sigma'_i = 0$, and let $\rho(\sigma, \sigma')$ denote the distance $\sum_{i=1}^m |\sigma_i - \sigma'_i|$. Then

$$R(F_\sigma, P'_{\sigma'}) = R(F_\sigma, P_\sigma) + \frac{\delta \Delta^2}{2m} \rho(\sigma, \sigma').$$

Moreover, for every k -points nearest neighbor quantizer F there exists σ and τ such that

$$\forall P_{\sigma(\tau')} \quad R(F, P_{\sigma(\tau')}) \geq R(F_\sigma, P_{\sigma(\tau')}) \geq \frac{1}{2} R(F_{\sigma(\tau)}, P_{\sigma(\tau')}).$$

The proof of Lemma 27 is a slight modification of that of [31, Proposition 4.2]. For the sake of completeness it is given in Section C.1 of the Appendix. Let $\hat{\mathbf{t}}$ be an empirically designed vector in $(\mathbb{R}^d)^{(k)}$, and recall that $Pd_{P,h,\mathbf{t}}^2$ is defined as $P \sum_{j=1}^k \left[\|\cdot - m(P_{t_j,h})\|^2 + v(P_{t_j,h}) \right] \mathbb{1}_{V_j}$. According to Lemma 27, we may write

$$\begin{aligned} \inf_{\hat{\mathbf{t}}} \sup_{P|\text{Supp}(P) \subset B(0,K)} \mathbb{E}P(d_{P,h,\hat{\mathbf{t}}}^2 - d_{P,h,k}^2) &\geq \inf_{\hat{\mathbf{t}}} \sup_{P_{\sigma(\tau')}} \mathbb{E}R(\hat{\mathbf{t}}) - R(F_{\sigma(\tau')}, P_{\sigma(\tau')}) \\ &\geq \frac{1}{2} \inf_{\hat{\tau}} \sup_{P_{\sigma(\tau')}} \mathbb{E}R(F_{\sigma(\hat{\tau})}, P_{\sigma(\tau')}) - R(F_{\sigma(\tau')}, P_{\sigma(\tau')}) \\ &\geq \frac{\delta \Delta^2}{2m} \inf_{\hat{\tau}} \sup_{P_{\sigma(\tau')}} \mathbb{E}\rho(\hat{\tau}, \tau'), \end{aligned} \quad (16)$$

where $\hat{\tau}$ denotes an empirically designed element of $\{-1, +1\}^{\frac{m}{2}}$. Let μ denote the measure $\sum_{i=1}^m (\delta_{z_i} + \delta_{z_i+w_i})$. For any distribution P and Q having densities with respect to μ we denote by $H^2(P, Q)$ their Hellinger distance.

Lemma 28.

Let τ and τ' in $\{-1, 1\}^{\frac{m}{2}}$ such that $\rho(\tau, \tau') = 2$. Then

$$H^2(P_{\sigma(\tau)}^{\otimes n}, P_{\sigma(\tau')}^{\otimes n}) \leq \frac{4n\delta^2}{m} := \alpha.$$

The proof of Lemma 28 is a slight modification of the proof of [31, Lemma 4.5], and is given in Section C.2 in the Appendix. A direct application of Assouad's Lemma (see, e.g., [40, Theorem 2.12]) entails that, for $\alpha \leq 2$,

$$\inf_{\hat{\tau}} \sup_{\tau \in \{-1, 1\}^{\frac{m}{2}}} \mathbb{E}\rho(\hat{\tau}, \tau) \geq \frac{m}{4} \left(1 - \sqrt{\alpha(1 - \alpha/4)} \right).$$

For $\delta = \frac{\sqrt{m}}{2\sqrt{n}}$, (16) yields $\inf_{\hat{\mathbf{t}}} \sup_{P|\text{Supp}(P) \subset B(0,K)} \mathbb{E}P(d_{P,h,\hat{\mathbf{t}}}^2 - d_{P,h,k}^2) \geq c_0 \frac{K^2 k^{\frac{1}{2} - \frac{2}{d}}}{\sqrt{n}}$. This proves (7).

Now denote by A the event $\bigcup_{i=1}^m \{P_n(\{z_i\}) \leq h\} \cup \{P_n(\{z_i + w_i\}) \leq h\}$. If $\delta \leq \frac{1}{9}$ and $h \leq \frac{1}{2k}$, using $\frac{(1-\delta)}{2m} - h \geq \frac{1}{6k}$ and a union of bounded difference inequalities (see, e.g., [11, Theorem 6.2]) leads to $\mathbb{P}_\sigma(A) \leq 2me^{-\frac{n}{72k^2}}$, where \mathbb{P}_σ denotes probability when the distribution of X_1, \dots, X_n is P_σ . Now, for $t \in \mathbb{R}^d$, for any σ , we may choose $P_{t,h} = \delta_z$, for some $z \in \{z_i, z_i + w_i\}_{i \in [1, m]}$, so that $m(P_{t,h}) = z$ and $v(P_{t,h}) = 0$.

Next, on the event A^c , we may choose for any σ and $t \in \mathbb{R}^d$ $P_{n,t,h} = P_{t,h}$. Thus $d_{P_n, h, \mathbf{t}}^2 \mathbb{1}_{A^c} = d_{P, h, \mathbf{t}}^2 \mathbb{1}_{A^c}$, where, for any $\mathbf{t} \in (\mathbb{R}^d)^{(k)}$, $d_{P, h, \mathbf{t}}^2$ is the function $\min_{j \in [1, k]} \|\cdot - m(P_{t_j, h})\|^2 + v(P_{t_j, h})$. Therefore, since for every σ , $\text{Supp}(P_\sigma) \subset B(0, K)$, we may write

$$\begin{aligned} \inf_{\hat{\mathbf{t}}} \sup_{\sigma} \mathbb{E}_\sigma P(d_{P_n, h, \hat{\mathbf{t}}}^2 - d_{P, h, k}^2) &\geq \inf_{\hat{\mathbf{t}}} \sup_{\sigma} \mathbb{E}_\sigma P(d_{P_n, h, \hat{\mathbf{t}}}^2 - d_{P, h, k}^2) \mathbb{1}_{A^c} - 16K^2 m e^{-\frac{n}{72k^2}} \\ &\geq \inf_{\hat{\mathbf{t}}} \sup_{\sigma} \mathbb{E}_\sigma P(d_{P, h, \hat{\mathbf{t}}}^2 - d_{P, h, k}^2) \mathbb{1}_{A^c} - 16K^2 m e^{-\frac{n}{72k^2}} \\ &\geq \inf_{\hat{\mathbf{t}}} \sup_{\sigma} \mathbb{E}_\sigma P(d_{P, h, \hat{\mathbf{t}}}^2 - d_{P, h, k}^2) - 32K^2 m e^{-\frac{n}{72k^2}}. \end{aligned}$$

Since $n \geq 14k$, $\delta = \frac{\sqrt{m}}{2\sqrt{n}} \leq \frac{1}{9}$ and (16) leads to the result again. \square

Appendix

For ease of understanding, the statements of the results are recalled before the proofs.

A Proofs of Lemmas 4, 23 and 24

A.1 Proof of Lemma 4

Let $P = \mathcal{U}_{S(0,1)}$, the uniform distribution on the circle, $Q_\beta = \beta \mathcal{U}_{S(0,1)} + (1 - \beta) \mathcal{U}_{B(0,1)}$ and τ^* be a minimizer of the k -means criterion $Q_\beta \min_{j \in [1,k]} \|\cdot - \tau_j\|^2$ (an optimal k -points codebook for Q_β).

Lemma (4).

Let d_{τ^*} denote the distance to τ^* function. Then, for k large enough,

$$\sup_{x \in S(0,1)} |d_{\tau^*}(x) - d_{S(0,1)}(x)| \leq C \left(\frac{1}{k^2} + (1 - \beta) \right)^{\frac{1}{3}}.$$

for some constant $C > 0$. On the other hand, for every $\rho > 0$, there exists $k_{\rho,\beta}$ such that, for all $k \geq k_{\rho,\beta}$, τ^* has at least one codepoint in $B(0, \rho)$.

Proof of Lemma 4. Let τ^* denote an optimal k -points codebook for Q_β and τ_P^* an optimal k -points codebook for P . We intend to bound $\sup_{x \in S(0,1)} |d_{S(0,1)}(x) - d_{\tau^*}(x)| = \sup_{x \in S(0,1)} d_{\tau^*}(x) = \Delta_{\infty, S(0,1)}$. Let $\Delta_P^2 = Pd_{\tau_P^*}^2$. We get that

$$\begin{aligned} \Delta_P^2 &\leq 2(Q_\beta d_{\tau^*}^2 + W_2^2(P, Q_\beta)) \\ &\leq 2(Q_\beta d_{\tau_P^*}^2 + (1 - \beta)) \\ &\leq 4Pd_{\tau_P^*}^2 + 6(1 - \beta). \end{aligned}$$

Comparing with a uniform grid on $S(0, 1)$, it is immediate that $Pd_{\tau_P^*}^2 \leq \frac{C}{k^2}$. Thus

$$\Delta_P^2 \leq C \left(\frac{1}{k^2} + (1 - \beta) \right).$$

Let $x \in S(0, 1)$ achieving the maximum distance $d_{\tau^*}(x) = \sup_{y \in S(0,1)} d_{\tau^*}(y)$. Since d_{τ^*} is 1-Lipschitz, we deduce that

$$B\left(x, \frac{\Delta_{\infty, S(0,1)}}{2}\right) \subset \left\{y \mid d_{\tau^*}(y) \geq \frac{\Delta_{\infty, S(0,1)}}{2}\right\}.$$

Since $P\left(B\left(x, \frac{\Delta_{\infty, S(0,1)}}{2}\right)\right) \geq C\left(\frac{\Delta_{\infty, S(0,1)}}{2}\right) \wedge 1$, Markov inequality yields that

$$\Delta_P^2 \geq C \left(\frac{\Delta_{\infty, S(0,1)}}{2} \right)^3 \wedge \left(\frac{\Delta_{\infty, S(0,1)}}{2} \right)^2.$$

Thus we have

$$\sup_{x \in S(0,1)} |d_{\tau^*}(x) - d_{S(0,1)}(x)| = \Delta_{\infty, S(0,1)} \leq C \Delta_P^{\frac{2}{3}} \vee \Delta_P.$$

Therefore, for k large enough,

$$\Delta_{\infty, S(0,1)} \leq C \left(\frac{1}{k^2} + (1 - \beta) \right)^{\frac{1}{3}}.$$

Now, assume that for all $j \in \llbracket 1, k \rrbracket$, $\tau_j^* \notin \overline{B}(0, \rho)$. Then

$$Q_\beta \min_{j \in \llbracket 1, k \rrbracket} \|\cdot - \tau_j^*\|^2 \geq (1 - \beta) \left(\frac{\rho}{2}\right)^4.$$

Let τ be a k -points codebook and a $\frac{\sqrt{2}}{\sqrt{k}}$ -covering of $B(0, 1)$ (a uniform grid for instance). We have

$$(1 - \beta) \left(\frac{\rho}{2}\right)^4 \leq Q_\beta \min_{j \in \llbracket 1, k \rrbracket} \|\cdot - \tau_j^*\|^2 \leq Q_\beta \min_{j \in \llbracket 1, k \rrbracket} \|\cdot - \tau_j\|^2 \leq \frac{2}{k},$$

hence the contradiction for k large enough. \square

A.2 Proof of Lemma 23

Lemma 23 follows from basic properties of sub-Gaussian distributions, that are recalled below. We recall that a sub-Gaussian measure Q with variance $V^2 > 0$ is such that $Q(B(0, t)^c) \leq \exp(-\frac{t^2}{2V^2})$ for all $t > V$ and that the set of sub-Gaussian measures is denoted by $\mathcal{P}^{(V)}(\mathbb{R}^d)$.

Lemma 29.

Let $Q \in \mathcal{P}^{(V)}(\mathbb{R}^d)$, a sub-Gaussian measure with variance $V^2 > 0$. Then, the following bounds are satisfied:

$$Q\|\cdot\| \leq 3V, \quad Q\|\cdot\|^2 \leq 3V^2.$$

As a consequence, for all $n \in \mathbb{N}^*$, with $R = \sqrt{4V^2(1+p)\log(n)}$ we have:

$$Q\mathbb{1}_{\|\cdot\| \geq R} \leq n^{-2p-2}, \quad Q(\|\cdot\| \mathbb{1}_{\|\cdot\| \geq R}) \leq \sqrt{3}Vn^{-(1+p)}, \quad Q(\|\cdot\|^2 \mathbb{1}_{\|\cdot\| \geq R}) \leq 3V^2n^{-(1+p)},$$

as well as

$$\mathbb{P}\left(\sup_{i \in \llbracket 1, n \rrbracket} \|X_i\| \geq R\right) \leq n^{-2p-1},$$

where $\mathbb{X}_n = \{X_1, X_2, \dots, X_n\}$ is an n -sample from Q .

At last, for $Q_h \in \mathcal{P}_h(Q)$ (that is such that hQ_h is a submeasure of Q with Q -mass h), we have

$$Q_h\|\cdot\|^2 \leq \frac{3V^2}{h}, \quad \|Q_h\cdot\| \leq \frac{\sqrt{3}V}{\sqrt{h}}.$$

Proof of Lemma 29. For the first inequality, note that

$$\begin{aligned} Q\|\cdot\| &= \int_{u \in \mathbb{R}^d} \int_{t=0}^{+\infty} \mathbb{1}_{t \leq \|u\|} dt Q(du) \\ &= \int_{t=0}^V Q(\|\cdot\| \geq t) dt + \int_{t=V}^{+\infty} Q(\|\cdot\| \geq t) dt \\ &\leq V + \int_{t=V}^{+\infty} e^{-\frac{t^2}{2V^2}} dt. \end{aligned}$$

Moreover, for X distributed according to the standard normal distribution $\mathcal{N}(0, 1)$, we have

$$\begin{aligned} \int_{t=V}^{+\infty} \exp\left(-\frac{t^2}{2V^2}\right) dt &= \sqrt{2\pi}V \mathbb{P}(X \geq 1) \\ &\leq \sqrt{2\pi}V \frac{\mathbb{E}[\exp(\lambda X)]}{\exp(\lambda)} \\ &= \sqrt{2\pi}V \exp\left(-\lambda + \frac{\lambda^2}{2}\right). \end{aligned}$$

The choice $\lambda = 1$ yields that

$$Q\|\cdot\| \leq \left(1 + \sqrt{2\pi} \exp\left(-\frac{1}{2}\right)\right) V.$$

For the second inequality, note that

$$\begin{aligned} Q\|\cdot\|^2 &= Q\|\cdot\|^2 \mathbb{1}_{\|\cdot\| \leq V} + Q\|\cdot\|^2 \mathbb{1}_{\|\cdot\| > V} \\ &\leq V^2 + Q\|\cdot\|^2 \mathbb{1}_{\|\cdot\| > V}. \end{aligned}$$

Then, we get

$$\begin{aligned} Q\|\cdot\|^2 \mathbb{1}_{\|\cdot\| > V} &= \int_{u \in \mathbb{R}^d} \mathbb{1}_{\|u\| > V} \int_{t=0}^{+\infty} 2t \mathbb{1}_{t \leq \|u\|} dt Q(du) \\ &= \int_{t=0}^V 2t Q(\|\cdot\| > V) dt + \int_{t=V}^{+\infty} 2t Q(\|\cdot\| \geq t) dt \\ &\leq \int_{t=0}^V 2te^{-\frac{1}{2}} dt + \int_{t=V}^{+\infty} 2te^{-\frac{t^2}{2V^2}} dt \\ &= 3V^2 e^{-\frac{1}{2}}. \end{aligned}$$

Thus, we have

$$Q\|\cdot\|^2 \mathbb{1}_{\|\cdot\| > V} \leq 2V^2, \quad (17)$$

and,

$$Q\|\cdot\|^2 \leq 3V^2.$$

The bound on $\sup_{i \in [1, n]} \|X_i\|$ derives from

$$\mathbb{P}\left(\sup_{i \in [1, n]} \|X_i\| \geq R\right) \leq n\mathbb{P}(\|X_1\| \geq R) \leq n^{-2p-1}.$$

Let $Q_h \in \mathcal{P}_h(Q)$. We may write

$$\begin{aligned} Q\|\cdot\|^2 &\leq \frac{1}{h} Q\|\cdot\|^2 \\ &\leq \frac{1}{h} [Q\|\cdot\|^2 \mathbb{1}_{\|\cdot\| \leq V} + Q\|\cdot\|^2 \mathbb{1}_{\|\cdot\| > V}] \\ &\leq \frac{V^2}{h} + \frac{Q\|\cdot\|^2 \mathbb{1}_{\|\cdot\| > V}}{h} \\ &\leq 3\frac{V^2}{h}. \end{aligned}$$

The last inequality follows from (17). \square

A.3 Proof of Lemma 24

Lemma (24).

If Y is a random variable sampled from a distribution P in $\mathcal{P}^K(\mathbb{R}^d)$ and Z is independent from Y and sampled from a distribution Q' in $\mathcal{P}^{(\sigma)}(\mathbb{R}^d)$ for some $\sigma > 0$. Then, the distribution Q of the random variable $X = Y + Z$ is sub-Gaussian with variance $V^2 = (K + \sigma)^2$, that is in $\mathcal{P}^{(K+\sigma)}(\mathbb{R}^d)$.

Moreover,

$$W_1(P, Q) \leq 3\sigma \text{ and } W_2(P, Q) \leq \sqrt{3}\sigma.$$

Proof of Lemma 24. For all $z \geq \sigma$ it holds

$$\begin{aligned}\mathbb{P}(\|Y + Z\| \geq K + z) &\leq \mathbb{P}(\|Z\| \geq z) \\ &\leq \exp -\frac{z^2}{2\sigma^2} \\ &\leq \exp -\frac{(z + K)^2}{2(\sigma + K)^2}.\end{aligned}$$

The L_1 -Wasserstein distance between P and Q satisfies

$$W_1(P, Q) = \mathbb{E} [\|(Y + Z) - Y\|] = Q' \|\cdot\|,$$

that is bounded from above by 3σ according to Lemma 29. The L_2 -Wasserstein distance between P and Q satisfies

$$W_2(P, Q) = \sqrt{\mathbb{E} [\|(Y + Z) - Y\|^2]} = \sqrt{Q' \|\cdot\|^2},$$

which is bounded from above by $\sqrt{3}\sigma$ according to Lemma 29. □

B Deviation inequalities for the proof of Theorem 19

This section is devoted to the proofs of Lemmas 25 and 26. We first expose the main technical tools from empirical processes theory that will be needed.

B.1 Preliminary results and tools to bound suprema of empirical processes

In this section, \mathcal{F} is a countable family of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$. We recall results that allow to derive upper bounds for the maximal deviation $\sup_{f \in \mathcal{F}} |Pf - P_n f|$. Most of these results can be found in [11, 10]. Their actual form is to be found in [12, Section 3.A].

We first recall the so-called bounded difference inequality [11, Theorem 6.2].

Theorem 30.

Assume that the elements of \mathcal{F} take values in $[-R, R]$. Then for every $\delta > 0$, with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}} |Pf - P_n f| \leq \mathbb{E} \left[\sup_{f \in \mathcal{F}} |Pf - P_n f| \right] + \sqrt{\frac{2R^2}{n} \log \left(\frac{1}{\delta} \right)}.$$

When the elements of \mathcal{F} are non-negative, the right-hand term can be replaced with $\sqrt{\frac{R^2}{2n} \log \left(\frac{1}{\delta} \right)}$.

Next, we will need tools to derive bounds on the expectation of suprema of empirical processes. We begin by recalling the Vapnik-Chervonenkis inequality [10, Theorem 3.4], for the particular case of indicator functions.

Theorem 31.

For all countable family \mathcal{F} of indicator functions of some collection \mathcal{A} , it holds

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} |Pf - P_n f| \right] \leq 2 \sqrt{\frac{2VC(\mathcal{A}) \log(n+1)}{n}}$$

and, for some universal constant C ,

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} |Pf - P_n f| \right] \leq C \sqrt{\frac{VC(\mathcal{A})}{n}},$$

where $VC(\mathcal{A})$ denotes the Vapnik-Chervonenkis dimension of the collection \mathcal{A} (see, e.g., [11, Example 3.10]).

At last, we will need a more general result to bound the expectation of the maximal deviation between P and P_n , over an arbitrary family of functions \mathcal{F} .

Theorem 32.

Assume that \mathcal{F} contains the null function and that $|f| \leq R$, for all $f \in \mathcal{F}$. Then

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} |Pf - P_n f| \right] \leq 24 \frac{R}{\sqrt{n}} \mathbb{E} \left[\int_0^{\frac{1}{2}} \sqrt{\log \left(N'_{\|\cdot\|} \left(\frac{u}{2}, \frac{(\mathcal{F} \cup -\mathcal{F})(X_1^n)}{R\sqrt{n}} \right) \right)} du \right],$$

where $(\mathcal{F} \cup -\mathcal{F})(X_1^n)$ denotes the set $\{(f(X_1), f(X_2), \dots, f(X_n)) \mid f \in \mathcal{F} \text{ or } -f \in \mathcal{F}\}$, and, for $S \subset \mathbb{R}^d$, $\delta \geq 0$, $N'_{\|\cdot\|}(\delta, S)$ denotes the δ -covering number of S with respect to the norm $\|\cdot\|$.

Theorem 32 is a particular instance of the so-called Dudley's entropy integral [11, Corollary 13.2]. In what follows, $\|\cdot\|$ will denote the Euclidean norm on \mathbb{R}^n . Combining Theorem 30 and Theorem 32 results in deviation bounds, provided that the covering number of the targeted class of functions is controlled. This is the aim of the following section.

B.2 Tools for covering numbers

We first recall some basic combinatorial properties of covering numbers. If \mathcal{F}_1 and \mathcal{F}_2 are two sets of measurable functions, we let $\mathcal{F}_1 + \mathcal{F}_2 = \{f_1 + f_2 \mid f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$ and $\mathcal{F}_1 \times \mathcal{F}_2 = \{f_1, f_2 \mid f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$.

Lemma 33.

For $\delta \geq 0$,

$$N'_{\|\cdot\|}(2\delta, \mathcal{F}_1 + \mathcal{F}_2(x_1^n)) \leq N'_{\|\cdot\|}(\delta, \mathcal{F}_1(x_1^n))N'_{\|\cdot\|}(\delta, \mathcal{F}_2(x_1^n)). \quad (18)$$

In addition, if the elements of \mathcal{F}_1 and \mathcal{F}_2 are all bounded by 1, then we have

$$N'_{\|\cdot\|}(2\delta, \mathcal{F}_1 \times \mathcal{F}_2(x_1^n)) \leq N'_{\|\cdot\|}(\delta, \mathcal{F}_1(x_1^n))N'_{\|\cdot\|}(\delta, \mathcal{F}_2(x_1^n)). \quad (19)$$

For a class of measurable functions \mathcal{F} , we denote by $-\mathcal{F} = \{-f \mid f \in \mathcal{F}\}$, and $\mathcal{F}_{(k)} = \{\min_{l \in \llbracket 1, k \rrbracket} f_l \mid f_l \in \mathcal{F}\}$. We have

$$N'_{\|\cdot\|}(\delta, \mathcal{F} \cup -\mathcal{F}(x_1^n)) \leq 2N'_{\|\cdot\|}(\delta, \mathcal{F}(x_1^n)), \quad (20)$$

as well as

$$N'_{\|\cdot\|}(\delta, \mathcal{F}_{(k)}(x_1^n)) \leq \left(N'_{\|\cdot\|}(\delta, \mathcal{F}(x_1^n))\right)^k. \quad (21)$$

Proof of Lemma 33. For the three first inequalities, if $\bigcup_{i \in \llbracket 1, N' \rrbracket} B_{\|\cdot\|}(c^i, \delta)$ covers $\mathcal{F}_1(x_1^n)$ and $\bigcup_{i \in \llbracket 1, M' \rrbracket} B_{\|\cdot\|}(c'^i, \delta)$ covers $\mathcal{F}_2(x_1^n)$, then it suffices to take balls centered at elements $c^i + c'^j$, $c^i c'^j$ or respectively c^i and $-c^i$ to cover $\mathcal{F}_1 + \mathcal{F}_2$, $\mathcal{F}_1 \times \mathcal{F}_2$ and respectively $\mathcal{F}_1 \cup -\mathcal{F}_1$.

For the fourth inequality, choose $c^1, c^2, \dots, c^{N'}$ in \mathbb{R}^n that satisfy the inclusion $\bigcup_{j \in \llbracket 1, N' \rrbracket} B_{\|\cdot\|}(c^j, \delta) \supset \mathcal{F}(x_1^n)$. For $f_1, f_2, \dots, f_k \in \mathcal{F}$, we let $c^{(1)}, c^{(2)}, \dots, c^{(k)}$ in \mathbb{R}^n be such that $f_j(x_1^n) \in B_{\|\cdot\|}(c^{(j)}, \delta)$ for all $j \in \llbracket 1, k \rrbracket$. Then, for all $i \in \llbracket 1, n \rrbracket$, we have

$$\left(\min_{l \in \llbracket 1, k \rrbracket} f_l(x_i) - \min_{l \in \llbracket 1, k \rrbracket} c^{(l)}(x_i) \right)^2 \leq \max_{l \in \llbracket 1, k \rrbracket} \left(f_l(x_i) - c^{(l)}(x_i) \right)^2.$$

Also note that the number of elements of the type $\min_{l \in \llbracket 1, k \rrbracket} c^{(l)}(x_1^n)$ is at most $\binom{N'}{k} \leq N'^k$. Equation (21) follows. \square

It is possible to relate the notion of covering number of a set $\mathcal{F}(x_1^n)$ to the notion of *shattering dimension* of the family \mathcal{F} .

Definition 34.

Given some $\delta > 0$, a subset S of \mathbb{R}^d is δ -shattered by \mathcal{F} if there exists some function h on S such that

$$\forall S' \subset S, \exists f \in \mathcal{F}, \text{ s.t. } \forall x \in S', f(x) \leq h(x) \text{ and } \forall x \in S \setminus S', f(x) \geq h(x) + \delta.$$

As for the Vapnik-Chervonenkis dimension, the shattering dimension of \mathcal{F} is defined as the maximal cardinality of a set S that is δ -shattered by \mathcal{F} , and is denoted by $V(\mathcal{F}, c\delta)$.

It is immediate that, for any $\delta \geq 0$, $V(\mathcal{F}, \delta) \leq VC(\mathcal{F})$. The following result allows to bound the δ -covering number of a set $\mathcal{F}(x_1^n)$ in terms of the $c\delta$ -shattering dimension of \mathcal{F} , $V(\mathcal{F}, c\delta)$.

Theorem 35.

[35, Theorem 1] *Let \mathcal{F} be a class of functions bounded by 1. There exist absolute constants K and c such that, for all $x_1^n \in (\mathbb{R}^d)^n$ and $\delta \in (0, 1)$,*

$$N'_{\|\cdot\|} \left(\delta, \frac{1}{\sqrt{n}} \mathcal{F}(x_1^n) \right) \leq \left(\frac{2}{\delta} \right)^{KV(\mathcal{F}, c\delta)}.$$

Thus, controlling shattering dimensions is enough to bound covering numbers. At last, we expose two lemmas that will be useful to control the shattering dimensions involved in the proofs of Lemmas 25 and 26.

Lemma 36.

Let $\mathcal{F} = \{ \mathbb{1}_{B(x,r)} \mid x \in \mathbb{R}^d, r > 0 \}$ for some $R > 0$. Then, for all $\delta \in (0, 1)$,

$$V(\mathcal{F}, \delta) \leq d + 1.$$

Proof of Lemma 36. According to [21], $VC(\mathcal{F}) \leq d + 1$. Thus, for any $\delta \in (0, 1)$, $V(\mathcal{F}, \delta) \leq VC(\mathcal{F}) \leq d + 1$. \square

Lemma 37.

For $R > 0$, let $\mathcal{F} = \{ x \mapsto \frac{1}{R} \mathbb{1}_{B(0,R)}(x) \langle x, v \rangle \mid v \in S(0, 1) \}$. Then, for all $\delta \in (0, 1)$,

$$V(\mathcal{F}, \delta) \leq d + 2.$$

Proof of Lemma 37. Let $S = \{x_1, x_2, \dots, x_V\}$ denote a subset of $B(0, R)$ that is δ -shattered by \mathcal{F} , and let h denote the function defined in Definition 34. Then for all $S' \subset S$, there is some $v \in S(0, 1)$ such that $\langle (x, -h(x)), (v, 1) \rangle \leq 0$ if $x \in S'$ and $\langle (x, -h(x)), (v, 1) \rangle \geq \delta$ if $x \in S \setminus S'$. In particular, all subsets $(S', -h(S'))$ rewrite as the intersection of $(S, -h(S))$ with a half space in \mathbb{R}^{d+1} . According to [21], this is possible only if $|S| \leq d + 2$. Hence, $V(\mathcal{F}, \delta) \leq d + 2$. \square

B.3 Application to truncated moments

As a direct consequence of the bounded inequality theorem together with the Vapnik-Chervonenkis inequality, it is possible to derive bounds for empirical processes of indicators of balls.

Corollary 38.

Let $P \in \mathcal{P}(\mathbb{R}^d)$. Then, there exists an absolute constant $C > 0$ such that, for all $p > 0$, with probability larger than $1 - 2n^{-p}$,

$$\sup_{x \in \mathbb{R}^d, r > 0} |P(\bar{B}(x, r)) - P_n(\bar{B}(x, r))| \leq C \sqrt{\frac{d+1}{n}} + \sqrt{\frac{p \log n}{2n}},$$

$$\sup_{x \in \mathbb{R}^d, r > 0} |P(B(x, r)) - P_n(B(x, r))| \leq C \sqrt{\frac{d+1}{n}} + \sqrt{\frac{p \log n}{2n}}.$$

In particular,

$$\sup_{x \in \mathbb{R}^d, r > 0} |P(\partial B(x, r)) - P_n(\partial B(x, r))| \leq 2C \sqrt{\frac{d+1}{n}} + 2\sqrt{\frac{p \log n}{2n}}.$$

Proof of Corollary 38. Corollary 38 follows from Theorem 30 with $\delta = n^{-p}$, combined with Theorem 32, Theorem 35 and Lemma 36 for the expectation part. \square

Next, we can bound the deviations of truncated moments of order 1, for the particular case of sub-Gaussian distributions.

Corollary 39.

If $P \in \mathcal{P}^{(V)}(\mathbb{R}^d)$ is a sub-Gaussian measure for some $V > 0$, then for all $p > 0$, with probability larger than $1 - 3n^{-p}$, it holds

$$\sup_{x \in \mathbb{R}^d, r > 0} \|(P_n - P) \cdot \mathbb{1}_{B(x,r)}\| \leq CV\sqrt{d+2} \frac{(1+p)\log(n)}{\sqrt{n}},$$

and

$$\sup_{x \in \mathbb{R}^d, r > 0} \|(P_n - P) \cdot \mathbb{1}_{\bar{B}(x,r)}\| \leq CV\sqrt{d+2} \frac{(1+p)\log(n)}{\sqrt{n}}.$$

In particular,

$$\sup_{x \in \mathbb{R}^d, r > 0} \|(P_n - P) \cdot \mathbb{1}_{\partial B(x,r)}\| \leq 2CV\sqrt{d+2} \frac{(1+p)\log(n)}{\sqrt{n}}.$$

Proof of Corollary 39. Let $\lambda = p \log(n)$, and $R = \sqrt{4V^2(\log(n) + \lambda)}$. We may write

$$\begin{aligned} \sup_{x,r} \|(P_n - P) \cdot \mathbb{1}_{B(x,r)}\| &= \sup_{x,r} \left\| \frac{1}{n} \sum_{i=1}^n X_i \mathbb{1}_{B(x,r)}(X_i) - P \cdot \mathbb{1}_{B(x,r)} \right\| \\ &\leq \sup_{x,r} \left\| \frac{1}{n} \sum_{i=1}^n X_i \mathbb{1}_{B(x,r)}(X_i) \mathbb{1}_{\|X_i\| \leq R} - P \cdot \mathbb{1}_{B(x,r)} \mathbb{1}_{\|\cdot\| \leq R} \right\| \\ &\quad + P \cdot \|\mathbb{1}_{\|\cdot\| > R}\| + \frac{1}{n} \sum_{i=1}^n \|X_i\| \mathbb{1}_{\|X_i\| > R}. \end{aligned}$$

From Lemma 29, we get

$$P(\|\cdot\| \mathbb{1}_{\|\cdot\| > R}) \leq 2Vn^{-(p+1)}.$$

Thus, with probability larger than $1 - n^{-2p-1}$, it holds

$$\frac{1}{n} \sum_{i=1}^n \|X_i\| \mathbb{1}_{\|X_i\| > R} = 0.$$

Next, using Cauchy-Schwarz inequality, we may write

$$\sup_{x \in \mathbb{R}^d, r > 0} \|(P_n - P) \cdot \mathbb{1}_{B(x,r)} \mathbb{1}_{B(0,R)}\| = \sup_{x \in \mathbb{R}^d, r > 0, v \in S(0,1)} |(P_n - P)f_{x,r,v}|,$$

where $f_{x,r,v}(u) = \langle v, u \mathbb{1}_{B(x,r)}(u) \mathbb{1}_{B(0,R)}(u) \rangle$ for all $u \in \mathbb{R}^d$.

Choosing $\delta = n^{-p}$ in Theorem 30 yields that, with probability larger than $1 - n^{-p}$,

$$\sup_{x \in \mathbb{R}^d, r > 0, v \in S(0,1)} |P_n - P|f_{x,r,v} \leq \mathbb{E} \left[\sup_{x \in \mathbb{R}^d, r > 0, v \in S(0,1)} |P_n - P|f_{x,r,v} \right] + \sqrt{\frac{2R^2 p \log(n)}{n}}.$$

It remains to derive an upper bound for $\mathbb{E} \left[\sup_{x \in \mathbb{R}^d, r > 0, v \in S(0,1)} |P_n - P|f_{x,r,v} \right]$. According to Theorem 32, we have

$$\mathbb{E} \left[\sup_{x \in \mathbb{R}^d, r > 0, v \in S(0,1)} |P_n - P|f_{x,r,v} \right] \leq 24 \frac{R}{\sqrt{n}} \mathbb{E} \left[\int_0^{\frac{1}{2}} \sqrt{\log \left(N'_{\|\cdot\|} \left(\frac{u}{2}, \frac{1}{R\sqrt{n}} \mathcal{F}(X_1^n) \right) \right)} du \right],$$

where $\frac{1}{R}\mathcal{F}(X_1^n)$ denotes $\mathcal{F}_1 \times \mathcal{F}_2(X_1^n)$, for \mathcal{F}_1 the set of indicators of balls and \mathcal{F}_2 the set of functions $x \mapsto \frac{1}{R}\mathbb{1}_{B(0,R)}\langle x, v \rangle$, for any $v \in S(0,1)$. According to Lemma 33, (19), for all $\delta > 0$,

$$\begin{aligned} N'_{\|\cdot\|} \left(\frac{u}{2}, \frac{1}{\sqrt{n}}\mathcal{F}_1 \times \mathcal{F}_2(x_1^n) \right) &= N'_{\|\cdot\|} \left(\frac{u}{2}\sqrt{n}, \mathcal{F}_1 \times \mathcal{F}_2(x_1^n) \right) \\ &\leq N'_{\|\cdot\|} \left(\frac{u\sqrt{n}}{4}, \mathcal{F}_1(x_1^n) \right) N'_{\|\cdot\|} \left(\frac{u\sqrt{n}}{4}, \mathcal{F}_2(x_1^n) \right) \\ &\leq N'_{\|\cdot\|} \left(\frac{u}{4}, \frac{1}{\sqrt{n}}\mathcal{F}_1(x_1^n) \right) N'_{\|\cdot\|} \left(\frac{u}{4}, \frac{1}{\sqrt{n}}\mathcal{F}_2(x_1^n) \right). \end{aligned}$$

According to Lemma 36, $V(\mathcal{F}_1, \delta) \leq d+1$. Thus, Theorem 35 entails

$$N'_{\|\cdot\|} \left(\frac{u}{4}, \frac{1}{\sqrt{n}}\mathcal{F}_1(x_1^n) \right) \leq \left(\frac{8}{u} \right)^{K(d+1)}.$$

From Lemma 37, $V(\mathcal{F}_2, \delta) \leq d+2$, thus

$$N'_{\|\cdot\|} \left(\frac{u}{4}, \frac{1}{\sqrt{n}}\mathcal{F}_2(x_1^n) \right) \leq \left(\frac{8}{u} \right)^{K(d+2)}.$$

As a consequence,

$$N'_{\|\cdot\|} \left(\frac{u}{2}, \frac{1}{R\sqrt{n}}\mathcal{F}(x_1^n) \right) \leq \left(\frac{8}{u} \right)^{K(2d+3)}.$$

Gathering the above inequalities yields that, with probability larger than $1 - 2n^{-p}$,

$$\begin{aligned} \sup_{x \in \mathbb{R}^d, r > 0, v \in S(0,1)} |P_n - P|f_{x,r,v} &\leq CR \frac{\sqrt{2d+3}}{\sqrt{n}} + 2Vn^{-(p+1)} + \sqrt{\frac{2R^2p \log(n)}{n}} \\ &\leq CV\sqrt{d+2} \frac{(1+p) \log(n)}{\sqrt{n}}, \end{aligned}$$

for some absolute positive constant C . □

At last, we derive an upper-bound for the deviation of moments of order 2 restricted to balls as follows.

Corollary 40.

Let $Q \in \mathcal{P}^{(V)}(\mathbb{R}^d)$ be a sub-Gaussian measure on \mathbb{R}^d with variance V^2 . Then, with probability larger than $1 - 3n^{-p}$,

$$\begin{aligned} \sup_{x \in \mathbb{R}^d, r > 0} |Q \cdot \|\cdot\|^2 \mathbb{1}_{B(x,r)} - Q_n \cdot \|\cdot\|^2 \mathbb{1}_{B(x,r)}| &\leq CV^2\sqrt{d+1} \frac{((1+p) \log(n))^{\frac{3}{2}}}{\sqrt{n}}, \\ \sup_{x \in \mathbb{R}^d, r > 0} |Q \cdot \|\cdot\|^2 \mathbb{1}_{\bar{B}(x,r)} - Q_n \cdot \|\cdot\|^2 \mathbb{1}_{\bar{B}(x,r)}| &\leq CV^2\sqrt{d+1} \frac{((1+p) \log(n))^{\frac{3}{2}}}{\sqrt{n}}. \end{aligned}$$

In particular,

$$\sup_{x \in \mathbb{R}^d, r > 0} |Q \cdot \|\cdot\|^2 \mathbb{1}_{\partial B(x,r)} - Q_n \cdot \|\cdot\|^2 \mathbb{1}_{\partial B(x,r)}| \leq 2CV^2\sqrt{d+1} \frac{((1+p) \log(n))^{\frac{3}{2}}}{\sqrt{n}}.$$

Proof of Corollary 40. Let $\lambda = p \log(n)$ and $R = \sqrt{4V^2(\log(n) + \lambda)}$. We may write

$$\begin{aligned} & \sup_{x,r} |(Q_n - Q)\|\cdot\|^2 \mathbb{1}_{B(x,r)}| \\ &= \sup_{x,r} \left| \frac{1}{n} \sum_{i=1}^n \|X_i\|^2 \mathbb{1}_{B(x,r)}(X_i) - Q(\|\cdot\|^2 \mathbb{1}_{B(x,r)}) \right| \\ &\leq \sup_{x,r} \left| \frac{1}{n} \sum_{i=1}^n \|X_i\|^2 \mathbb{1}_{B(x,r)}(X_i) \mathbb{1}_{\|X_i\| \leq R} - Q(\|\cdot\|^2 \mathbb{1}_{B(x,r)} \mathbb{1}_{\|\cdot\| \leq R}) \right| \\ &\quad + Q(\|\cdot\|^2 \mathbb{1}_{\|\cdot\| > R}) + \frac{1}{n} \sum_{i=1}^n \|X_i\|^2 \mathbb{1}_{\|X_i\| > R}. \end{aligned}$$

Using Lemma 29, we get

$$Q(\|\cdot\|^2 \mathbb{1}_{\|\cdot\| > R}) \leq 3V^2 n^{-(p+1)},$$

and, with probability larger than $1 - n^{-1-2p}$,

$$\frac{1}{n} \sum_{i=1}^n \|X_i\|^2 \mathbb{1}_{\|X_i\| > R} = 0.$$

Thus, choosing $\delta = n^{-p}$ in Theorem 30 yields that, with probability larger than $1 - n^{-p}$,

$$\begin{aligned} & \sup_{x,r} \left| \frac{1}{n} \sum_{i=1}^n \|X_i\|^2 \mathbb{1}_{B(x,r)}(X_i) \mathbb{1}_{B(0,R)}(X_i) - Q(\|\cdot\|^2 \mathbb{1}_{B(x,r)} \mathbb{1}_{B(0,R)}) \right| \leq \\ & \mathbb{E} \left[\sup_{x,r} \left| \frac{1}{n} \sum_{i=1}^n \|X_i\|^2 \mathbb{1}_{B(x,r)}(X_i) \mathbb{1}_{B(0,R)}(X_i) - Q(\|\cdot\|^2 \mathbb{1}_{B(x,r)} \mathbb{1}_{B(0,R)}) \right| \right] \\ & \quad + \sqrt{\frac{R^4 p \log(n)}{2n}}. \end{aligned}$$

Applying Theorem 32 gives

$$\begin{aligned} & \mathbb{E} \left[\sup_{x,r} \left| \frac{1}{n} \sum_{i=1}^n \|X_i\|^2 \mathbb{1}_{B(x,r)}(X_i) \mathbb{1}_{B(0,R)}(X_i) - Q(\|\cdot\|^2 \mathbb{1}_{B(x,r)} \mathbb{1}_{B(0,R)}) \right| \right] \\ & \leq 24 \frac{R^2}{\sqrt{n}} \mathbb{E} \left[\int_0^{\frac{1}{2}} \sqrt{\log \left(N'_{\|\cdot\|} \left(\frac{u}{2}, \frac{1}{R^2 \sqrt{n}} (\mathcal{F} \cup -\mathcal{F})(X_1^n) \right) \right)} du \right], \end{aligned}$$

where $\frac{1}{R^2} (\mathcal{F} \cup -\mathcal{F})(X_1^n) = (\mathcal{F}_1 \cup -\mathcal{F}_1) \times \mathcal{F}_2(X_1^n)$, and \mathcal{F}_1 is the set of indicators of balls and \mathcal{F}_2 is the single function $\{x \mapsto \frac{\|x\|^2}{R^2} \mathbb{1}_{B(0,R)(x)}\}$. According to Lemma 33, (19), for all $\delta > 0$,

$$N'_{\|\cdot\|} \left(2\delta, \frac{1}{R^2 \sqrt{n}} (\mathcal{F}_1 \cup -\mathcal{F}_1) \times \mathcal{F}_2(x_1^n) \right) \leq N'_{\|\cdot\|} \left(\delta, \frac{1}{\sqrt{n}} (\mathcal{F}_1 \cup -\mathcal{F}_1)(x_1^n) \right) N'_{\|\cdot\|} \left(\delta, \frac{1}{\sqrt{n}} \mathcal{F}_2(x_1^n) \right).$$

Using Lemma 33, (20), we get

$$N'_{\|\cdot\|} \left(\delta, \frac{1}{\sqrt{n}} (\mathcal{F}_1 \cup -\mathcal{F}_1)(x_1^n) \right) \leq 2N'_{\|\cdot\|} \left(\delta, \frac{1}{\sqrt{n}} \mathcal{F}_1(x_1^n) \right).$$

Recall that Lemma 36 entails $V(\mathcal{F}_1, \delta) \leq d + 1$. Thus, combining the above inequalities with Theorem 35 gives

$$N'_{\|\cdot\|} \left(\delta, \frac{1}{\sqrt{n}} (\mathcal{F} \cup -\mathcal{F})(x_1^n) \right) \leq 2 \left(\frac{2}{\delta} \right)^{K(d+1)},$$

for some positive constant K . At last, gathering the above results yields that, with probability larger than $1 - 2n^{-p}$,

$$\begin{aligned} & \sup_{x \in \mathbb{R}^d, r > 0} |Q| \cdot \|\cdot\|^2 \mathbb{1}_{B(x,r)} - Q_n \|\cdot\|^2 \mathbb{1}_{B(x,r)} \\ & \leq \sqrt{d+1} \frac{R^2}{\sqrt{n}} C + R^2 \sqrt{\frac{p \log(n)}{2n}} + 3V^2 n^{-p-1} \\ & \leq CV^2 \sqrt{d+1} \frac{((1+p) \log(n))^{\frac{3}{2}}}{\sqrt{n}}, \end{aligned}$$

for some absolute positive constant C . \square

B.4 Proof of Lemma 26

Lemma (26).

Assume that Q is sub-Gaussian with variance V^2 , then, for every $p > 0$, with probability larger than $1 - 9n^{-p}$, we have

$$\begin{aligned} \sup_{t \in \mathbb{R}^d} \|m(Q_{t,h}) - m(Q_{n t,h})\| & \leq \frac{CV \sqrt{d}(p+1) \log(n)}{h \sqrt{n}}, \\ \sup_{t \in (\mathbb{R}^d)^{(k)}} |Q_n(\gamma - \hat{\gamma})(t, \cdot)| & \leq CV^2 \frac{\sqrt{d}(p+1)^{\frac{3}{2}} \log(n)^{\frac{3}{2}}}{h \sqrt{n}}. \end{aligned}$$

The proof of Lemma 26 is based on the following deviation bounds.

Lemma 41.

Suppose that $Q \in \mathcal{P}^{(V)}(\mathbb{R}^d)$ for some $V > 0$. Then, for every $p > 0$, with probability larger than $1 - 4n^{-p}$, we have,

$$\begin{aligned} \sup_{x,r} |(Q_n - Q)| \mathbb{1}_{B(x,r)} & \leq C \sqrt{d+1} \frac{\sqrt{p \log(n)}}{\sqrt{n}} \\ \sup_{x,r} \|(Q_n - Q) \cdot \mathbb{1}_{B(x,r)}\| & \leq CV \sqrt{d+2} \frac{(p+1) \log(n)}{\sqrt{n}} \\ \sup_{x,r} |(Q_n - Q)| \cdot \|\cdot\|^2 \mathbb{1}_{B(x,r)} & \leq CV^2 \sqrt{d+1} \frac{((p+1) \log(n))^{\frac{3}{2}}}{\sqrt{n}} \end{aligned}$$

where $C > 0$ denotes a universal constant, $x \in \mathbb{R}^d$, $v \in S(0,1)$, $r > 0$ and $c \in \mathbb{R}$.

Proof of Lemma 41. This is a direct consequence of the proofs of Corollary 39, 40 and 38. \square

We are now in position to prove Lemma 26.

Proof of Lemma 26. Let $t \in \mathbb{R}^d$, and denote by $r = \delta_{Q,h}(t)$, $r_n = \delta_{Q_n,h}(t)$, and $R = 2V \sqrt{(p+1) \log(n)}$. We may write, for $Q_{t,h} = \frac{1}{h} Q \mathbb{1}_{B(t,r)} + \frac{\alpha}{h} Q \mathbb{1}_{\partial B(t,r)}$ and $Q_{n t,h} =$

$\frac{1}{h}Q_n \mathbb{1}_{B(t, r_n)} + \frac{\alpha_n}{h}Q \mathbb{1}_{\partial B(t, r_n)}$ for non-negative α and α_n ,

$$\begin{aligned}
\|m(Q_{t,h}) - m(Q_{n t,h})\| &\leq \frac{1}{h}(\|Q \cdot \mathbb{1}_{B(t,r)} + \alpha Q \cdot \mathbb{1}_{\partial B(t,r)} - Q \cdot \mathbb{1}_{B(t,r_n)} - \alpha_n Q \cdot \mathbb{1}_{\partial B(t,r_n)}\| \\
&\quad + \|Q \cdot \mathbb{1}_{B(t,r_n)} - Q_n \cdot \mathbb{1}_{B(t,r_n)}\| + \alpha_n \|Q \cdot \mathbb{1}_{\partial B(t,r_n)} - Q_n \cdot \mathbb{1}_{\partial B(t,r_n)}\|) \\
&\leq \frac{1}{h}(Q \| \cdot \| |\mathbb{1}_{B(t,r)} + \alpha \mathbb{1}_{\partial B(t,r)} - \mathbb{1}_{B(t,r_n)} - \alpha_n \mathbb{1}_{\partial B(t,r_n)}| + \|(Q - Q_n) \cdot \mathbb{1}_{B(t,r_n)}\| \\
&\quad + \|(Q - Q_n) \cdot \mathbb{1}_{\partial B(t,r_n)}\|) \\
&\leq \frac{1}{h}(RQ |\mathbb{1}_{B(t,r)} + \alpha \mathbb{1}_{\partial B(t,r)} - \mathbb{1}_{B(t,r_n)} - \alpha_n \mathbb{1}_{\partial B(t,r_n)}| + \|(Q - Q_n) \cdot \mathbb{1}_{B(t,r_n)}\| \\
&\quad + \|(Q - Q_n) \cdot \mathbb{1}_{\partial B(t,r_n)}\| + Q \| \cdot \| \mathbb{1}_{\|\cdot\| > R}).
\end{aligned}$$

Moreover, considering the case $r > r_n$, we have

$$\begin{aligned}
&Q |\mathbb{1}_{B(t,r)} + \alpha \mathbb{1}_{\partial B(t,r)} - \mathbb{1}_{B(t,r_n)} - \alpha_n \mathbb{1}_{\partial B(t,r_n)}| \\
&= Q(B(t,r)) + \alpha Q(\partial B(t,r)) - Q(B(t,r_n)) - \alpha_n Q(\partial B(t,r_n)) \\
&= h - Q(B(t,r_n)) - \alpha_n Q(\partial B(t,r_n)) \\
&\leq |(Q_n - Q) \mathbb{1}_{B(t,r_n)} + \alpha_n \mathbb{1}_{\partial B(t,r_n)}| \\
&\leq |Q_n - Q| \mathbb{1}_{B(t,r_n)} + |Q_n - Q| \mathbb{1}_{\partial B(t,r_n)}.
\end{aligned}$$

The same inequality holds when $r \leq r_n$. On the event described in Lemma 41, using Lemma 29, we have that

$$\begin{aligned}
\|(Q - Q_n) \cdot \mathbb{1}_{B(t,r_n)}\| &\leq CV \sqrt{d} \frac{(p+1) \log(n)}{\sqrt{n}}, \\
\|(Q - Q_n) \cdot \mathbb{1}_{\partial B(t,r_n)}\| &\leq CV \sqrt{d} \frac{(p+1) \log(n)}{\sqrt{n}}, \\
|Q_n(B(t,r_n)) - Q(B(t,r_n))| &\leq C \sqrt{d} \frac{\sqrt{(p+1) \log(n)}}{\sqrt{n}}, \\
|Q_n(\partial B(t,r_n)) - Q(\partial B(t,r_n))| &\leq C \sqrt{d} \frac{\sqrt{(p+1) \log(n)}}{\sqrt{n}}, \\
Q \| \cdot \| \mathbb{1}_{\|\cdot\| > R} &\leq 2V n^{-(p+1)}.
\end{aligned}$$

Thus, we get

$$\sup_{t \in \mathbb{R}^d} \|m(Q_{t,h}) - m(Q_{n t,h})\| \leq \frac{CV \sqrt{d} (p+1) \log(n)}{h \sqrt{n}}, \quad (22)$$

that is the first equation of Lemma 26. For $\mathbf{t} \in (\mathbb{R}^d)^k$, we get that

$$\begin{aligned}
|\gamma(\mathbf{t}, x) - \hat{\gamma}(\mathbf{t}, x)| &\leq \max_{j \in [1,k]} | -2 \langle x, m(Q_{t_j,h}) - m(Q_{n t_j,h}) \rangle + (M(Q_{t_j,h}) - M(Q_{n t_j,h})) | \\
&\leq 2 \|x\| \max_{j \in [1,k]} \|m(Q_{t_j,h}) - m(Q_{n t_j,h})\| + \max_{j \in [1,k]} |M(Q_{t_j,h}) - M(Q_{n t_j,h})|.
\end{aligned}$$

As well, $\sup_{t \in \mathbb{R}^d} |M(Q_{t,h}) - M(Q_{n t,h})|$ is bounded from above by

$$\begin{aligned}
&\frac{1}{h}(R^2 |Q_n - Q| \mathbb{1}_{B(t,r_n)} + R^2 |Q_n - Q| \mathbb{1}_{\partial B(t,r_n)} + \|(Q - Q_n) \cdot \mathbb{1}_{B(t,r_n)}\| \\
&\quad + \|(Q - Q_n) \cdot \mathbb{1}_{\partial B(t,r_n)}\| + Q \| \cdot \|^2 \mathbb{1}_{\|\cdot\| > R}).
\end{aligned}$$

Using Lemma 41 and 29 again yields

$$\begin{aligned}
|(Q - Q_n)\|\cdot\|^2 \mathbb{1}_{\mathbb{B}(t, r_n)}| &\leq CV^2 \sqrt{d} \frac{(p+1)^{\frac{3}{2}} \log(n)^{\frac{3}{2}}}{\sqrt{n}} \\
|(Q - Q_n)\|\cdot\|^2 \mathbb{1}_{\partial\mathbb{B}(t, r_n)}| &\leq CV^2 \sqrt{d} \frac{(p+1)^{\frac{3}{2}} \log(n)^{\frac{3}{2}}}{\sqrt{n}} \\
|(Q - Q_n)\mathbb{1}_{\mathbb{B}(t, r_n)}| &\leq C\sqrt{d} \frac{\sqrt{(p+1)\log(n)}}{\sqrt{n}} \\
|(Q - Q_n)\mathbb{1}_{\partial\mathbb{B}(t, r_n)}| &\leq C\sqrt{d} \frac{\sqrt{(p+1)\log(n)}}{\sqrt{n}} \\
Q\|\cdot\|^2 \mathbb{1}_{\|\cdot\| > R} &\leq 3V^2 n^{-(p+1)}.
\end{aligned}$$

Collecting all pieces leads to

$$|\gamma(\mathbf{t}, x) - \hat{\gamma}(\mathbf{t}, x)| \leq C\|x\| \frac{V\sqrt{d}(p+1)\log(n)}{h\sqrt{n}} + CV^2 \frac{\sqrt{d}(p+1)^{\frac{3}{2}} \log(n)^{\frac{3}{2}}}{h\sqrt{n}}. \quad (23)$$

At last, according to Lemma 29

$$\mathbb{P}\left\{\max_i \|X_i\| \geq R\right\} \leq n^{-2p-1}.$$

We deduce that

$$Q_n |\gamma(\mathbf{t}, \cdot) - \hat{\gamma}(\mathbf{t}, \cdot)| \leq CV^2 \frac{\sqrt{d}(p+1)^{\frac{3}{2}} \log(n)^{\frac{3}{2}}}{h\sqrt{n}},$$

with probability larger than $1 - 9n^{-p}$. \square

B.5 Proof of Lemma 25

Lemma (25).

If Q is sub-Gaussian with variance V^2 , then, for every $p > 0$, with probability larger than $1 - 2n^{-p}$, we have

$$\sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} |(Q - Q_n)\gamma(\mathbf{t}, \cdot)| \leq C \frac{V^2 \sqrt{kd}(1+p)^{\frac{3}{2}} \log(n)^{\frac{3}{2}}}{h\sqrt{n}},$$

for some absolute positive constant C .

Proof of Lemma 25. With the notation $l_{t_i}(x) = -2\langle x, m(Q_{t_i, h}) \rangle + \|m(Q_{t_i, h})\|^2 + v(Q_{t_i, h})$, we get that:

$$\sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} |(Q - Q_n)\gamma(\mathbf{t}, \cdot)| = \sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} \left| (Q - Q_n) \min_{i \in [1, k]} l_{t_i} \right|.$$

First we note that since Q is sub-Gaussian with variance V^2 , it follows from Lemma 29 that, for every $c \in \mathbb{R}^d$,

$$\|m(Q_{c, h})\|^2 + v(Q_{c, h}) = Q_{c, h}\|\cdot\|^2 \leq \frac{3V^2}{h}. \quad (24)$$

Set $R = 2V\sqrt{\log(n) + \lambda}$ and $\lambda = p\log(n)$. Then, according to Lemma 29, with probability larger than $1 - n^{-2p-1}$,

$$\max_{i \in [1, n]} \|X_i\| \leq R. \quad (25)$$

We may then write

$$\begin{aligned}
\sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} |(Q - Q_n)\gamma(\mathbf{t}, \cdot)| &= \sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} \left| \frac{1}{n} \sum_{i=1}^n \gamma(\mathbf{t}, X_i) - Q\gamma(\mathbf{t}, \cdot) \right| \\
&\leq \sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} \left| \frac{1}{n} \sum_{i=1}^n \gamma(\mathbf{t}, X_i) \mathbb{1}_{\|X_i\| \leq R} - Q\gamma(\mathbf{t}, \cdot) \mathbb{1}_{\|\cdot\| \leq R} \right| \\
&\quad + \sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} |Q\gamma(\mathbf{t}, \cdot) \mathbb{1}_{\|\cdot\| > R}| + \sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} \left| \frac{1}{n} \sum_{i=1}^n \gamma(\mathbf{t}, X_i) \mathbb{1}_{\|X_i\| > R} \right|.
\end{aligned}$$

According to (25), the last part is 0 with probability larger than $1 - n^{-2p-1}$. Moreover, according to Lemma 29,

$$\begin{aligned}
Q|\gamma(\mathbf{t}, \cdot)| \mathbb{1}_{\|\cdot\| > R} &\leq Q \mathbb{1}_{\|\cdot\| > R} \sup_{i \in \llbracket 1, k \rrbracket} 2|\langle \cdot, m(Q_{t_i, h}) \rangle| + \|m(Q_{t_i, h})\|^2 + v(Q_{t_i, h}) \\
&\leq 2\sqrt{3} \frac{V}{\sqrt{h}} Q \|\cdot\| \mathbb{1}_{\|\cdot\| > R} + \frac{3V^2}{h} Q \mathbb{1}_{\|\cdot\| > R} \\
&\leq \frac{3V^2}{h} n^{-1-p} (2\sqrt{h} + n^{-1-p}) \\
&\leq \frac{9V^2}{h} n^{-1-p}.
\end{aligned}$$

It remains to bound

$$\sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} |(Q - Q_n)\gamma(\mathbf{t}, \cdot) \mathbb{1}_{\|\cdot\| \leq R}|.$$

Since Lemma 29 ensures that, for every \mathbf{t} and u , $|\gamma(\mathbf{t}, u) \mathbb{1}_{\|u\| \leq R}| \leq \left(R + \frac{V\sqrt{3}}{\sqrt{h}}\right)^2 := Z$, then Theorem 30 entails

$$\begin{aligned}
\mathbb{P} \left(\sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} |(Q - Q_n)\gamma(\mathbf{t}, \cdot) \mathbb{1}_{\|\cdot\| \leq R}| \geq \mathbb{E} \sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} |(Q - Q_n)\gamma(\mathbf{t}, \cdot) \mathbb{1}_{\|\cdot\| \leq R}| + Z\sqrt{\frac{2\lambda}{n}} \right) \\
\leq e^{-\lambda} = n^{-p}.
\end{aligned}$$

Then, Theorem 32 yields

$$\begin{aligned}
\mathbb{E} \sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} |(Q - Q_n)\gamma(\mathbf{t}, \cdot) \mathbb{1}_{\|\cdot\| \leq R}| \\
\leq 24 \frac{Z}{\sqrt{n}} \mathbb{E} \left[\int_0^{\frac{1}{2}} \sqrt{\log \left(N'_{\|\cdot\|} \left(\frac{u}{2}, \frac{1}{Z\sqrt{n}} (\mathcal{F}_{(k)} \cup -\mathcal{F}_{(k)})(X_1^n) \right) \right)} du \right],
\end{aligned}$$

with $\frac{1}{Z\sqrt{n}} \mathcal{F}_{(k)} = \left\{ \frac{1}{Z\sqrt{n}} \mathbb{1}_{\|\cdot\| \leq R} \min_{i \in \llbracket 1, k \rrbracket} l_{t_i} \right\}$. According to (20), (21) and (18), it holds that

$$\begin{aligned}
N'_{\|\cdot\|} \left(\frac{u}{2}, \frac{1}{Z\sqrt{n}} (\mathcal{F}_{(k)} \cup -\mathcal{F}_{(k)})(X_1^n) \right) \\
\leq 2N'_{\|\cdot\|} \left(\frac{u}{2}, \frac{1}{Z\sqrt{n}} (\mathcal{F}_{(k)})(X_1^n) \right) \\
\leq 2 \left(N'_{\|\cdot\|} \left(\frac{u}{2}, \frac{1}{Z\sqrt{n}} (\mathcal{F})(X_1^n) \right) \right)^k \\
\leq 2 \left(N'_{\|\cdot\|} \left(\frac{u}{4}, \frac{1}{Z\sqrt{n}} (\mathcal{G}_1)(X_1^n) \right) \times N'_{\|\cdot\|} \left(\frac{u}{4}, \frac{1}{Z\sqrt{n}} (\mathcal{G}_2)(X_1^n) \right) \right)^k,
\end{aligned}$$

where $\mathcal{G}_1 = \left\{ x \mapsto \frac{-2\langle x, m(Q_{t,h}) \rangle \mathbb{1}_{\|x\| \leq R}}{Z\sqrt{n}} \mid t \in \mathbb{R}^d \right\}$ and \mathcal{G}_2 denotes the set $\mathcal{G}_2 = \left\{ x \mapsto \frac{(\|m(Q_{t,h})\|^2 + v(Q_{t,h})) \mathbb{1}_{\|x\| \leq R}}{Z\sqrt{n}} \mid t \in \mathbb{R}^d \right\}$. According to Theorem 35 and Lemma 37, we have

$$N'_{\|\cdot\|}(\delta, \mathcal{G}_1(X_1^n)) \leq \left(\frac{2}{\delta} \right)^{V'(d+2)},$$

for some absolute positive constant V' . Also, $N'_{\|\cdot\|}(\delta, \mathcal{G}_2(X_1^n)) \leq N'_{\|\cdot\|}(\delta, \mathcal{G}_3(X_1^n)) \leq \frac{2}{\delta}$ with $\mathcal{G}_3 = \left\{ x \mapsto \frac{t \mathbb{1}_{\|x\| \leq R}}{\sqrt{n}} \mid t \in [0, 1] \right\}$. As a consequence, with probability larger than $1 - 2n^{-p}$, it holds

$$\begin{aligned} \sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} |(Q - Q_n)\gamma(\mathbf{t}, \cdot)| &\leq CZ \frac{\sqrt{k}\sqrt{d+2}}{\sqrt{n}} + \frac{9V^2}{h} n^{-1-p} + Z \sqrt{\frac{2p \log(n)}{n}} \\ &\leq C \frac{V^2 \sqrt{kd}(1+p)^{\frac{3}{2}} \log(n)^{\frac{3}{2}}}{h\sqrt{n}}, \end{aligned}$$

for some positive absolute constant C . □

C Proofs of intermediate Lemmas for Proposition 21

This section is devoted to the proofs of Lemmas 27 and 28 that are needed to prove Proposition 21.

We recall that $m = 2k/3$, z_1, \dots, z_m is a 6Δ -net in $B(0, K)$, with $\Delta = K/(6m^{\frac{1}{d}})$, so that such a net exists ; w_1, \dots, w_m in \mathbb{R}^d are such that $\|w_i\| = \Delta$ and $z_i + w_i \in B(0, K)$. For $\sigma \in \{-1, +1\}^m$ such that $\sum_{i=1}^m \sigma_i = 0$, P_σ denotes the distribution that satisfies, for $i \in \llbracket 1, m \rrbracket$,

$$P_\sigma(\{z_i\}) = P_\sigma(\{z_i + w_i\}) = \frac{(1 + \sigma_i \delta)}{2m},$$

with $\delta \leq \frac{1}{3}$. For $\tau \in \{-1, 1\}^{\frac{m}{2}}$, $\sigma(\tau)$ is defined by $\sigma(\tau)_j = \tau_j$ and $\sigma(\tau)_{\frac{m}{2}+j} = -\tau_j$, for $j \in \llbracket 1, m/2 \rrbracket$. A p -points quantizer F is a map from \mathbb{R}^d to \mathbb{R}^d such that $|F(\mathbb{R}^d)| = p$, and F_σ is the k -points quantizer satisfying

$$\begin{aligned} F_\sigma(z_i) &= z_i, & F_\sigma(z_i + w_i) &= z_i + w_i & \text{if } & \sigma_i = +1 \\ F_\sigma(z_i) &= F_\sigma(z_i + w_i) = z_i & & & \text{if } & \sigma_i = -1. \end{aligned}$$

At last, for a quantizer F with images q_1, \dots, q_p and sets of preimages V_1, \dots, V_p , $R(F, P_\sigma)$ denotes the quantity

$$R(F, P_\sigma) = \sum_{i=1}^p P_\sigma \left[\|\cdot - m(P_{q_i, h})\|^2 + v(P_{q_i, h}) \right] \mathbb{1}_{V_i},$$

where for $i \in \llbracket 1, p \rrbracket$, $P_{q_i, h} \in \mathcal{P}_{q_i, h}(P_\sigma)$.

C.1 Proof of Lemma 27

Lemma (27).

Assume that $\delta \leq \frac{1}{3}$ and $h \leq \frac{1}{3m}$. Let σ and σ' be such that $\sum_{i=1}^m \sigma_i = \sum_{i=1}^m \sigma'_i = 0$, and let $\rho(\sigma, \sigma')$ denote the distance $\sum_{i=1}^m |\sigma_i - \sigma'_i|$. Then

$$R(F_\sigma, P'_\sigma) = R(F_\sigma, P_\sigma) + \frac{\delta \Delta^2}{2m} \rho(\sigma, \sigma').$$

Moreover, for every k -points nearest-neighbor quantizer F there exists σ and τ such that

$$\forall P_{\sigma(\tau')} \quad R(F, P_{\sigma(\tau')}) \geq R(F_\sigma, P_{\sigma(\tau')}) \geq \frac{1}{2} R(F_{\sigma(\tau)}, P_{\sigma(\tau')}).$$

Proof of Lemma 27. First recall that if $\delta \leq \frac{1}{3}$ and $h \leq \frac{1}{3m}$, then, for any σ and $t \in \mathbb{R}^d$, we may choose $P_{t, h} = \delta_z$, for some $z \in \{z_i, z_i + w_i\}_{i \in \llbracket 1, m \rrbracket}$, so that $m(P_{t, h}) = z$ and $v(P_{t, h}) = 0$. Thus, we may restrict our attention to nearest-neighbor quantizers F whose preimages are (non-weighted) Voronoi cells, with ties arbitrarily broken. In particular, we have that

$$R(F_\sigma, P'_\sigma) = R(F_\sigma, P_\sigma) + \frac{\delta \Delta^2}{2m} \rho(\sigma, \sigma').$$

Let F be such a nearest-neighbor quantizer with k codepoints, and for short denote by $R(F)$ the quantity $R(F, P_{\sigma(\tau')})$. The following construction provides F_σ (not depending on $\sigma(\tau')$) such that $R(F_\sigma) \leq R(F)$. Let V_i denote the union of the Voronoi cells associated

with z_i and $z_i + \omega_i$, in the Voronoi diagram generated by the sequences z and ω . We adopt the following notation.

$$\begin{cases} n_i(F) &= |F(\mathcal{B}(0, K)) \cap V_i|, \\ n_i^{out}(F) &= |F(V_i) \cap V_i^c|, \\ I_j(F) &= \{i \mid n_i(F) = j\}, \\ i_j(F) &= |I_j(F)|, \\ i_{\geq j}(F) &= \sum_{i \geq j} i_j(F). \end{cases}$$

The first step is to add codepoints to empty cells. From the k -quantizer F , a quantizer F_1 is built as follows.

- If $n_i(F) \geq 1$, then we take $F_{1|V_i} \equiv F|_{V_i}$.
- If $n_i(F) = 0$, then we set $F_1(V_i) = z_i$.

Notice that F_1 is a $(k + i_0(F))$ -quantizer. Let us denote $k_1 = k + i_0$ and $p_{\pm} = \frac{1 \pm \delta}{2m}$. Then $R(F_1)$ can be bounded as follows.

Let i be an integer between 1 and m . We denote by $R_i(F)$ the contribution of V_i to the risk $R(F)$. If $i \in I_{\geq 1}$, then $R_i(F) = R_i(F_1)$. Otherwise, if $i \in I_0(F)$,

$$R_i(F_1) = p_{\pm} \Delta^2.$$

Furthermore, if $i \in I_0$, then $n_i^{out}(F) \geq 1$. Without loss of generality assume that $F(z_i) \notin V_i$. Then $\|m(P_{F(z_i, h)} - z_i)\| \geq \frac{6\Delta - \Delta}{2}$. Since F is a nearest-neighbor quantizer and z_1, \dots, z_m is a 6Δ -net, we also have $F(z_i + w_i) \notin V_i$. Thus $\|m(P_{F(z_i + w_i, h)} - z_i + w_i)\| \geq \frac{6\delta - 2\Delta}{2}$. Hence

$$R_i(F) \geq p_{\pm} 8\Delta^2.$$

It follows that

$$R(F_1) \leq R(F) - 7i_0(F)p_- \Delta^2.$$

At last, we build the k -points quantizer F_{σ} according to the following rule: we choose $i_0(F)$ cells with $n_i(F) = 2$ and withdraw one point in every cell. Since $i_2(F) = i_0(F) + \frac{k}{3}$, $i_2(F) \geq i_0(F)$, thus such a procedure is valid. Elementary calculation show that

$$R(F_{\sigma}) \leq R(F) + i_0(F)(p_+ - 7p_-)\Delta^2 \leq R(F),$$

since $p_+ \leq 2p_-$.

Now let τ be in $\{-1, 1\}^{\frac{m}{2}}$ such that $\rho(\sigma(\tau), \sigma) = \min_{\tau'} \rho(\sigma(\tau'), \sigma)$. Recalling that $R(F_{\sigma}) = R(F_{\sigma}, P_{\sigma(\tau)})$, it follows that

$$\begin{aligned} R(F_{\sigma(\tau)}, P_{\sigma(\tau)}) &= R(F_{\sigma(\tau')}, P_{\sigma(\tau')}) + \frac{(p_+ - p_-)\Delta^2}{2} \rho(\sigma(\tau), \sigma(\tau')) \\ &\leq 2R(F_{\sigma(\tau')}, P_{\sigma(\tau')}) + \frac{(p_+ - p_-)\Delta^2}{2} (\rho(\sigma(\tau), \sigma) + \rho(\sigma, \sigma(\tau'))) \\ &\leq 2R(F_{\sigma}, P_{\sigma}) + 2\frac{(p_+ - p_-)\Delta^2}{2} \rho(\sigma(\tau'), \sigma) \\ &\leq 2R(F_{\sigma}, P_{\sigma(\tau')}). \end{aligned}$$

□

C.2 Proof of Lemma 28

Lemma (28).

Let τ and τ' in $\{-1, 1\}^{\frac{m}{2}}$ such that $\rho(\tau, \tau') = 2$. Then

$$H^2 \left(P_{\sigma(\tau)}^{\otimes n}, P_{\sigma(\tau')}^{\otimes n} \right) \leq \frac{4n\delta^2}{m} := \alpha.$$

Proof of Lemma 28. Let us introduce, for distributions P and Q with densities f and g with respect to the distribution $\mu = \sum_{i=1}^m (\delta_{z_i} + \delta_{z_i+w_i})$, the affinity

$$\alpha(P, Q) = \int \sqrt{fg} d\mu,$$

so that $H^2(P, Q) = 2(1 - \alpha(P, Q))$. Elementary calculation shows that, if $\rho(\sigma, \sigma') = 4$, then

$$\alpha(P_\sigma, P_{\sigma'}) = 1 + \frac{2}{m} \left(\sqrt{1 - \delta^2} - 1 \right) \geq 1 - \frac{2\delta^2}{m}.$$

Hence we deduce

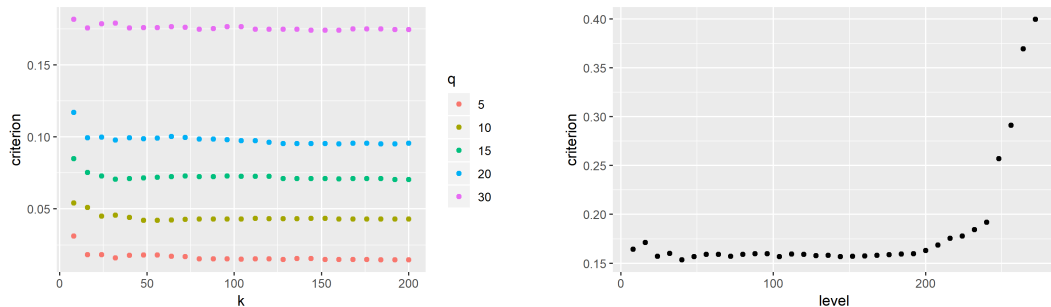
$$\begin{aligned} H^2(P_\sigma^{\otimes n}, P_{\sigma'}^{\otimes n}) &= 2(1 - \alpha(P_\sigma^{\otimes n}, P_{\sigma'}^{\otimes n})) \\ &= 2(1 - \alpha^n(P_\sigma, P_{\sigma'})) \\ &\leq \frac{4n\delta^2}{m}. \end{aligned}$$

Finally, we notice that $\rho(\tau, \tau') = 2$ implies $\rho(\sigma(\tau), \sigma(\tau')) = 4$, for τ, τ' in $\{-1, +1\}^{\frac{m}{2}}$. This gives the result. \square

D Additional figures

In this section, we give some more details about the numerical illustrations of Section 4.

D.1 Robust topological inference



Selection of the parameter k

Selection of the signal points number

Figure 4: Parameters selection

In the experiments, we have calibrated the mass parameter h (or equivalently the number of nearest-neighbors q) manually. We chose the largest q such the distance to the q -th nearest neighbor remains small with respect to the regularity of M , seized by the reach of M [25, Definition 4.1]. This heuristic resulted in choosing $q = 10$. According to the remark following Theorem 19, there is no need to take k as large as the sample size n for the k -PDTM to be a good approximation of d_M . This is confirmed by Figure 4 (left), where we plotted an approximation $\tilde{P}d_{\mathbb{X}_n, h, k}^2$ of the risk $Pd_{\mathbb{X}_n, h, k}^2$ as a function of k , for different values of q . Here, \tilde{P} is the uniform distribution on a set of points regularly generated on M and $d_{\mathbb{X}_n, h, k}$ the k -PDTM associated to the uniform distribution on \mathbb{X}_n . The curve strongly decreases and then stabilizes. We selected $k = 50$, the parameter k at the beginning of the stabilization phase.

D.2 Outlier detection

In this section, we give more details on the truncated versions of k -means, the q -witnessed distance and the k -PDTM. The proportion of signal points α can be calibrated from the data. In Figure 4 (right), we displayed the mean of the empirical k -PDTM for the points in its sublevel set that contains αn points, as a function of the level αn . We observe two jumps in the slope, one at $\alpha n = 200$ and the other one at $\alpha n = 240$. This makes sense since we generated 200 signal points, 40 outliers on a small rectangle and 40 others on a large rectangle.

We experimented the compression and denoising scheme with $\alpha n = 200$ for the three methods. Results are depicted by Figure 5. The points removed by the procedure are colored in pink. Note that for k -means, some points far from M are considered as signal points. On the contrary, we remove outliers and recover the features of M from the truncated versions of the k -PDTM and of the q -witnessed distance.

Note that similar graphics as Figure 4 (right) can be obtained for the trimmed k -PDTM procedure, so that it is possible to select the number of outliers from data in this case as well.

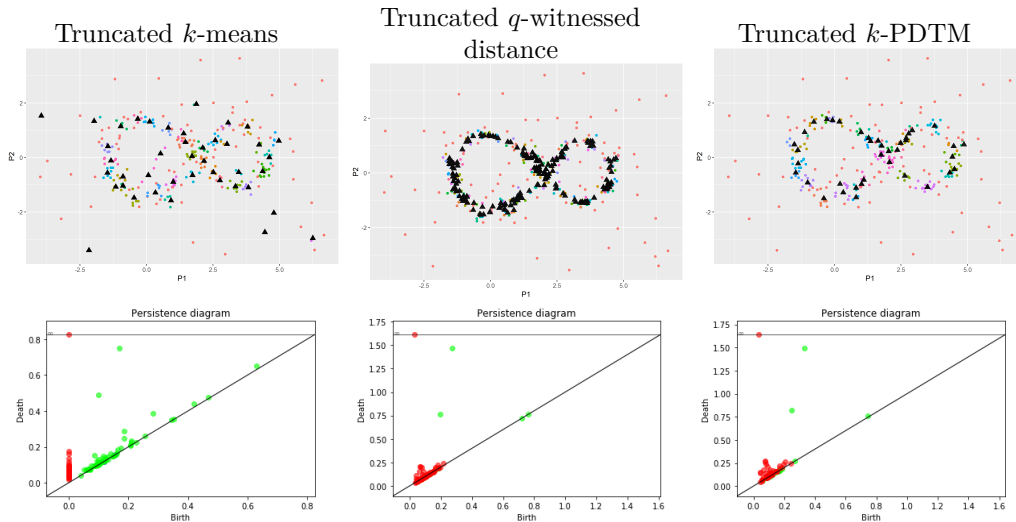


Figure 5: Comparison of the methods after thresholding

References

- [1] Eddie Aamari and Clément Levrard. “Nonasymptotic rates for manifold, tangent space and curvature estimation”. In: *Ann. Statist.* 47.1 (2019), pp. 177–204. ISSN: 0090-5364. DOI: 10.1214/18-AOS1685. URL: <https://doi.org/10.1214/18-AOS1685>.
- [2] Eddie Aamari and Clément Levrard. “Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction”. In: *Discrete Comput. Geom.* 59.4 (2018), pp. 923–971. ISSN: 0179-5376. DOI: 10.1007/s00454-017-9962-z. URL: <https://doi.org/10.1007/s00454-017-9962-z>.
- [3] Catherine Aaron and Alejandro Cholaquidis. “On boundary detection”. In: *ArXiv e-prints* (2019). arXiv: 1603.08460.
- [4] Charalambos D. Aliprantis and Kim C. Border. *Infinite dimensional analysis*. Third. A hitchhiker’s guide. Springer, Berlin, 2006, pp. xxii+703. ISBN: 978-3-540-32696-0; 3-540-32696-0.
- [5] Arindam Banerjee, Xin Guo, and Hui Wang. “On the optimality of conditional expectation as a Bregman predictor”. In: *IEEE Trans. Inform. Theory* 51.7 (2005), pp. 2664–2669. ISSN: 0018-9448. DOI: 10.1109/TIT.2005.850145. URL: <https://doi.org/10.1109/TIT.2005.850145>.
- [6] Arindam Banerjee et al. “Clustering with Bregman divergences”. In: *J. Mach. Learn. Res.* 6 (2005), pp. 1705–1749. ISSN: 1532-4435.
- [7] Peter L. Bartlett, Tamás Linder, and Gábor Lugosi. “The minimax distortion redundancy in empirical quantizer design”. In: *IEEE Trans. Inform. Theory* 44.5 (1998), pp. 1802–1813. ISSN: 0018-9448. DOI: 10.1109/18.705560. URL: <https://doi.org/10.1109/18.705560>.
- [8] Gérard Biau, Luc Devroye, and Gábor Lugosi. “On the performance of clustering in Hilbert spaces”. In: *IEEE Trans. Inform. Theory* 54.2 (2008), pp. 781–790. ISSN: 0018-9448. DOI: 10.1109/TIT.2007.913516. URL: <https://doi.org/10.1109/TIT.2007.913516>.

- [9] Jean-Daniel Boissonnat, Frédéric Chazal, and Mariette Yvinec. *Geometric and topological inference*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2018, pp. xii+233. ISBN: 978-1-108-41089-2; 978-1-108-41939-0. DOI: 10.1017/9781108297806. URL: <https://doi.org/10.1017/9781108297806>.
- [10] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. “Theory of classification: a survey of some recent advances”. In: *ESAIM Probab. Stat.* 9 (2005), pp. 323–375. ISSN: 1292-8100. DOI: 10.1051/ps:2005018. URL: <https://doi.org/10.1051/ps:2005018>.
- [11] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities. A nonasymptotic theory of independence*, With a foreword by Michel Ledoux. Oxford University Press, Oxford, 2013, pp. x+481. ISBN: 978-0-19-953525-5. DOI: 10.1093/acprof:oso/9780199535255.001.0001. URL: <http://dx.doi.org/10.1093/acprof:oso/9780199535255.001.0001>.
- [12] Claire BréchetEAU. “Vers une vision robuste de l’inférence géométrique”. PhD thesis. Université Paris Sud (Paris 11) - Université Paris Saclay, 2018.
- [13] Claire BréchetEAU, Aurélie Fischer, and Clément Levrard. “Robust Bregman clustering”. In: *ArXiv e-prints* (2018). arXiv: 1812.04356.
- [14] Mickaël Buchet et al. “Efficient and robust persistent homology for measures”. In: *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, Philadelphia, PA, 2015, pp. 168–180. DOI: 10.1137/1.9781611973730.13. URL: <https://doi.org/10.1137/1.9781611973730.13>.
- [15] Hervé Cardot, Peggy Cénac, and Pierre-André Zitt. “Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm”. In: *Bernoulli* 19.1 (2013), pp. 18–43. ISSN: 1350-7265. DOI: 10.3150/11-BEJ390. URL: <https://doi.org/10.3150/11-BEJ390>.
- [16] Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. “Geometric inference for probability measures”. In: *Found. Comput. Math.* 11.6 (2011), pp. 733–751. ISSN: 1615-3375. DOI: 10.1007/s10208-011-9098-0. URL: <https://doi.org/10.1007/s10208-011-9098-0>.
- [17] Frédéric Chazal, Pascal Massart, and Bertrand Michel. “Rates of convergence for robust geometric inference”. In: *Electron. J. Stat.* 10.2 (2016), pp. 2243–2286. ISSN: 1935-7524. DOI: 10.1214/16-EJS1161. URL: <https://doi.org/10.1214/16-EJS1161>.
- [18] Frédéric Chazal et al. “Convergence rates for persistence diagram estimation in topological data analysis”. In: *J. Mach. Learn. Res.* 16 (2015), pp. 3603–3635. ISSN: 1532-4435.
- [19] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. “Stability of persistence diagrams”. In: *Discrete Comput. Geom.* 37.1 (2007), pp. 103–120. ISSN: 0179-5376. DOI: 10.1007/s00454-006-1276-5. URL: <https://doi.org/10.1007/s00454-006-1276-5>.
- [20] J. A. Cuesta-Albertos, A. Gordaliza, and C. Matrán. “Trimmed k -means: an attempt to robustify quantizers”. In: *Ann. Statist.* 25.2 (1997), pp. 553–576. ISSN: 0090-5364. DOI: 10.1214/aos/1031833664. URL: <https://doi.org/10.1214/aos/1031833664>.
- [21] R. M. Dudley. “Balls in \mathbf{R}^k do not cut all subsets of $k+2$ points”. In: *Adv. in Math.* 31.3 (1979), pp. 306–308. ISSN: 0001-8708. DOI: 10.1016/0001-8708(79)90047-1. URL: [https://doi.org/10.1016/0001-8708\(79\)90047-1](https://doi.org/10.1016/0001-8708(79)90047-1).
- [22] Herbert Edelsbrunner. *Weighted Alpha Shapes*. Tech. rep. Champaign, IL, USA, 1992.

- [23] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. “Topological persistence and simplification”. In: *Discrete Comput. Geom.* 28.4 (2002). Discrete and computational geometry and graph drawing (Columbia, SC, 2001), pp. 511–533. ISSN: 0179-5376. DOI: 10.1007/s00454-002-2885-2. URL: <https://doi.org/10.1007/s00454-002-2885-2>.
- [24] Y. Eldar et al. “The Farthest Point Strategy for Progressive Image Sampling”. In: *Trans. Img. Proc.* 6.9 (1997), pp. 1305–1315. ISSN: 1057-7149. DOI: 10.1109/83.623193. URL: <https://doi.org/10.1109/83.623193>.
- [25] Herbert Federer. “Curvature measures”. In: *Trans. Amer. Math. Soc.* 93 (1959), pp. 418–491. ISSN: 0002-9947. DOI: 10.2307/1993504. URL: <https://doi.org/10.2307/1993504>.
- [26] Heinrich Fritz, Luis A. Garcia-Escudero, and Agustin Mayo-Iscar. “tclust: An R Package for a Trimming Approach to Cluster Analysis”. In: *Journal of Statistical Software* 47.12 (2012), pp. 1–26. URL: <http://www.jstatsoft.org/v47/i12/>.
- [27] Christopher R. Genovese et al. “Manifold estimation and singular deconvolution under Hausdorff loss”. In: *Ann. Statist.* 40.2 (2012), pp. 941–963. ISSN: 0090-5364. DOI: 10.1214/12-AOS994. URL: <https://doi.org/10.1214/12-AOS994>.
- [28] Leonidas Guibas, Dmitriy Morozov, and Quentin Mérigot. “Witnessed k -distance”. In: *Discrete Comput. Geom.* 49.1 (2013), pp. 22–45. ISSN: 0179-5376. DOI: 10.1007/s00454-012-9465-x. URL: <https://doi.org/10.1007/s00454-012-9465-x>.
- [29] Arlene K. H. Kim and Harrison H. Zhou. “Tight minimax rates for manifold estimation under Hausdorff loss”. In: *Electron. J. Stat.* 9.1 (2015), pp. 1562–1582. ISSN: 1935-7524. DOI: 10.1214/15-EJS1039. URL: <https://doi.org/10.1214/15-EJS1039>.
- [30] Hidetoshi Komiya. “Elementary proof for Sion’s minimax theorem”. In: *Kodai Math. J.* 11.1 (1988), pp. 5–7. ISSN: 0386-5991. DOI: 10.2996/kmj/1138038812. URL: <https://doi.org/10.2996/kmj/1138038812>.
- [31] Clément Levrard. “Nonasymptotic bounds for vector quantization in Hilbert spaces”. In: *Ann. Statist.* 43.2 (2015), pp. 592–619. ISSN: 0090-5364. DOI: 10.1214/14-AOS1293. URL: <https://doi.org/10.1214/14-AOS1293>.
- [32] Stuart P. Lloyd. “Least squares quantization in PCM”. In: *IEEE Trans. Inform. Theory* 28.2 (1982), pp. 129–137. ISSN: 0018-9448. DOI: 10.1109/TIT.1982.1056489. URL: <https://doi.org/10.1109/TIT.1982.1056489>.
- [33] Mauro Maggioni, Stanislav Minsker, and Nate Strawn. “Multiscale dictionary learning: non-asymptotic bounds and robustness”. In: *J. Mach. Learn. Res.* 17 (2016), Paper No. 2, 51. ISSN: 1532-4435.
- [34] Geoffrey McLachlan and David Peel. *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York, 2000, pp. xxii+419. ISBN: 0-471-00626-2. DOI: 10.1002/0471721182. URL: <https://doi.org/10.1002/0471721182>.
- [35] S. Mendelson and R. Vershynin. “Entropy and the combinatorial dimension”. In: *Invent. Math.* 152.1 (2003), pp. 37–55. ISSN: 0020-9910. DOI: 10.1007/s00222-002-0266-3. URL: <https://doi.org/10.1007/s00222-002-0266-3>.
- [36] Quentin Mérigot. “Lower bounds for k -distance approximation”. In: *Computational geometry (SoCG’13)*. ACM, New York, 2013, pp. 435–440. DOI: 10.1145/2462356.2462367. URL: <https://doi.org/10.1145/2462356.2462367>.

- [37] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. “Finding the homology of submanifolds with high confidence from random samples”. In: *Discrete Comput. Geom.* 39.1-3 (2008), pp. 419–441. ISSN: 0179-5376. DOI: 10.1007/s00454-008-9053-2. URL: <https://doi.org/10.1007/s00454-008-9053-2>.
- [38] Jeff M. Phillips, Bei Wang, and Yan Zheng. “Geometric inference on kernel density estimates”. In: *31st International Symposium on Computational Geometry*. Vol. 34. LIPIcs. Leibniz Int. Proc. Inform. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2015, pp. 857–871.
- [39] Alberto Rodríguez Casal. “Set estimation under convexity type assumptions”. In: *Ann. Inst. H. Poincaré Probab. Statist.* 43.6 (2007), pp. 763–774. ISSN: 0246-0203.
- [40] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. Springer, New York, 2009, pp. xii+214. ISBN: 978-0-387-79051-0. DOI: 10.1007/b13794. URL: <https://doi.org/10.1007/b13794>.
- [41] Afra Zomorodian and Gunnar Carlsson. “Computing persistent homology”. In: *Discrete Comput. Geom.* 33.2 (2005), pp. 249–274. ISSN: 0179-5376. DOI: 10.1007/s00454-004-1146-y. URL: <https://doi.org/10.1007/s00454-004-1146-y>.