The k-PDTM : a coreset for robust geometric inference

Claire Brécheteau, Clément Levrard

The k-PDTM : a coreset for robust geometric inference*

Brécheteau, Claire
claire.brecheteau@inria.fr
Université Paris-Saclay – LMO & Inria

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

January 27, 2018

Abstract

Analyzing the sub-level sets of the distance to a compact sub-manifold of $\mathbb{R}_d$ is a common method in TDA to understand its topology. The distance to measure (DTM) was introduced by Chazal, Cohen-Steiner and Mérigot in [7] to face the non-robustness of the distance to a compact set to noise and outliers. This function makes possible the inference of the topology of a compact subset of $\mathbb{R}_d$ from a noisy cloud of n points lying nearby in the Wasserstein sense. In practice, these sub-level sets may be computed using approximations of the DTM such as the q-witnessed distance [10] or other power distance [6]. These approaches lead eventually to compute the homology of unions of n growing balls, that might become intractable whenever n is large.

To simultaneously face the two problems of large number of points and noise, we introduce the k-power distance to measure (k-PDTM). This new approximation of the distance to measure may be thought of as a k-coreset based approximation of the DTM. Its sublevel sets consist in union of k-balls, $k << n$, and this distance is also proved robust to noise. We assess the quality of this approximation for k possibly dramatically smaller than n, for instance $k = n^{\frac{1}{3}}$ is proved to be optimal for 2-dimensional shapes. We also provide an algorithm to compute this k-PDTM.

Keywords : distance to a measure, geometric inference, coreset, power function, weighted Voronoï tesselation, empirical approximation

1 Introduction

1.1 Background on robust geometric inference

Let $M \subset \mathbb{R}^d$ be a compact set included in the closed Euclidean ball $\overline{B}(0, K)$, for $K > 0$, whose topology is to be inferred. A common approach is to sample $\mathbb{X}_n = \{X_1, X_2, \ldots, X_n\}$ on M, and approximate the distance to M via the distance to the sample points. As emphasized in [7], such an approach suffers from non-robustness to outliers. To face this issue, [7] introduces the *distance to measure* as a robust surrogate of the distance to M, when $\mathbb{X}_n = \{X_1, X_2, \ldots, X_n\}$ is considered as a n-sample, that is n independent realizations of a distribution measure P whose support $Supp(P)$ is M, possibly corrupted by noise. Namely, for a Borel probability measure P on $\mathbb{R}_d$, a mass

*This work was partially supported by the ANR project TopData and GUDHI

parameter $h \in [0,1]$ and $x \in \mathbb{R}_d$, the distance of x to the measure P, $\mathrm{d}_{P,h}(x)$ is defined by

Definition 1 (DTM)**.**

$$\mathrm{d}^2_{P,h}(x) = \frac{1}{h}\int_{l=0}^{h} \delta^2_{P,l}(x)\,\mathrm{d}l, \qquad \textit{with} \qquad \delta_{P,l}(x) = \inf\{r > 0 \mid P(\overline{\mathrm{B}}(x,r)) > l\},$$

where $\overline{\mathrm{B}}(x,r)$ denotes the closed Euclidean ball with radius r. When P is uniform enough on a compact set with positive reach ρ, this distance is proved to approximate well the distance to M ([7, Proposition 4.9]) and is robust to noise ([7, Theorem 3.5]). The distance to measure is usually inferred from $\mathbb{X}_n$ via its empirical counterpart, also 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.

Mérigot et al noted in [10] that 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 [10] (q-witnessed distance) and [6] (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-shape for instance; see [9]. Nonetheless, when n is large, there is still a need for a coreset allowing to efficiently compute an approximation of the DTM, as pointed out in [13]. In [12], Mérigot proves that such a coreset cannot be too small for large dimension.

1.2 Contribution

This paper aims at providing such a coreset for the DTM, to face the case where there are many observations, possibly corrupted by noise. We introduce the *k-power distance to a measure P* (k-PDTM), which is defined as the square root of one of the best k-power functions approximating the square of the DTM from above, for the $L_1(P)$ norm. Roughly, we intend to approximate the DTM of a point x with a power distance $\mathrm{d}_{P,h,k}(x)$ of the form

$$\mathrm{d}_{P,h,k}(x) = \sqrt{\min_{i \in [\![1,k]\!]} \|x - \theta_i\|^2 + \omega^2_{P,h}(\theta_i)},$$

where the θ_i's and corresponding ω's are suitably chosen. Its sub-level sets are union of k balls. Thus, the study of the associated topological invariants gets tractable in practice, even for massive data.

We begin by providing some theoretical guarantees on the k-PDTM we introduce. For instance, we prove that it can be expressed as a power distance from a coreset of k points that are local means of the measure P. The proofs rely on a geometric study of local sub-measures of P with fixed mass $h \in [0,1]$, showing that such a coreset makes sense whenever P is supported on a compact set. In particular, we prove that the set of means of local sub-measures of P is convex. The discrete case relies on the duality between a weighted Delaunay diagram and its associated weighted Voronoï diagram.

Once the k-PDTM properly defined, the main contribution of our paper are the following. First we assess that the k-DTM is a good approximation of the DTM in the L_1 sense (Proposition 18), showing for instance that whenever M has dimension d'

$$P\left(\mathrm{d}^2_{P,h,k}(u) - \mathrm{d}^2_{P,h}(u)\right) \leq C_{P,h} k^{-\frac{2}{d'}},$$

where $Pf(u)$ stands for the integration of f with respect to measure P. As mentioned in Proposition 22, this allows to infer topological guarantees from the sublevel sets of the k-PDTM.

Second we prove that this k-PDTM shares the robustness properties of the DTM with respect to Wasserstein deformations (Proposition 21). Namely, if Q is a sub-Gaussian deformation of P such that the Wasserstein distance $W_2(P,Q) \leq \sigma \leq K$, it holds

$$P\left|\mathrm{d}^2_{Q,h,k}(u) - \mathrm{d}^2_{P,h}(u)\right| \leq P\left(\mathrm{d}^2_{P,h,k}(u) - \mathrm{d}^2_{P,h}(u)\right) + C_{P,h}\sigma K,$$

ensuring that the approximation guarantees of our k-PDTM are stable with respect to Wasserstein noise. Similar to the DTM, this also guarantees that an empirical k-PDTM, that is built on $\mathbb{X}_n$, is a consistent approximation of the true k-PDTM.

At last, we provide more insights on the construction of the empirical k-PDTM from a point cloud $\mathbb{X}_n$, facing the practical situation where only a corrupted sample is at hand. We expose a k-means like algorithm with complexity $O(n^2hkd)$, and we analyze the approximation performance of such an empirical output. Theorem 24 shows that, with high probability,

$$P\left(\mathrm{d}^2_{P_n,h,k}(u) - \mathrm{d}^2_{P,h,k}(u)\right) \leq C_{P,h}\frac{\sqrt{k}\left(\log(n)\right)^{\frac{3}{2}}}{\sqrt{n}}.$$

Combining this estimation result with the approximation results between k-PDTM and DTM mentioned above suggest that an optimal choice for k is $k = n^{\frac{d'}{d'+4}}$, whenever M has dimension d', resulting in a deviation between empirical k-PDTM and DTM of order $n^{-1/(d'+4)}$. This has to be compared with the $n^{-1/d'}$ approximation that the empirical DTM achieves in such cases. In the case where n is large, this $n^{-1/(d'+4)}$ approximation suffices for topological inference. Thus, topological inference built on significantly less points might provide almost similar guarantees than the DTM.

1.3 Organization of the paper

This paper is organized as follows. In Section 2, we recall some definitions for the DTM that can be expressed as a power distance, and study the set of local means. Section 3 is devoted to the k-PDTM, a k-power distance which approximates the DTM. We make the link with two equivalent definitions for the k-PDTM, derive some stability results, prove its proximity to the DTM highlighting its interest for topological inference. The case of noisy point clouds is addressed in Section 4, where an algorithm to approximate the k-PDTM comes up with theoretical guarantees.

2 Some background about the DTM

2.1 Notation and definitions for the DTM

In the paper, we denote by $\mathbb{R}_d = \{x = (x_1, x_2, \ldots, x_d) \mid \forall i \in [\![1,d]\!],\, x_i \in \mathbb{R}\}$ the d-dimensional space equipped with the Euclidean norm $\|.\|$. For $k \in \mathbb{N}^*$ and any space $\mathcal{A}$, $\mathcal{A}^{(k)}$ stands for $\{t = (t_1, t_2, \ldots, t_k) \mid \forall i \in [\![1,k]\!],\, t_i \in \mathcal{A}\}$, where two elements are identified whenever they are equal up to a permutation of the coordinates. Also, $\mathrm{S}(0,r) = \{x \in \mathbb{R}_d \mid \|x\| = r\}$ denotes the Euclidean sphere of radius r, $\mathrm{B}(x,r) = \{y \in \mathbb{R}_d \mid \|x-y\| < r\}$ the Euclidean ball centred at x, and for $c \in \mathbb{R}$ and $v \in \mathrm{S}(0,1)$, $\mathrm{H}(v,c)$ denotes the half-space $\{x \in \mathbb{R}_d \mid \langle x, v\rangle \geq c\}$. Also, for any subset A of $\mathbb{R}_d$, $\overline{A}$ stands for its closure, A° for its interior, $\partial A = \overline{A} \backslash A^\circ$ its boundary and $A^c = \mathbb{R}_d \backslash A$ its complementary set in $\mathbb{R}_d$.

In the following, $\mathcal{P}(\mathbb{R}_d)$ stands for the set of Borel probability distributions P, with support $\mathrm{Supp}(P) \subset \mathbb{R}_d$, and, for any P-integrable function f, $Pf(u)$ denotes the expectation of f with respect to P. The following sets of distributions are of particular interest: we denote by $\mathcal{P}^K(\mathbb{R}_d) = \left\{P \in \mathcal{P}(\mathbb{R}_d) \mid \mathrm{Supp}(P) \subset \overline{\mathrm{B}}(0, K)\right\}$ for $K > 0$, and $\mathcal{P}^{K,h}(\mathbb{R}_d)$ is the set of $P \in \mathcal{P}^K(\mathbb{R}_d)$ which put mass neither on the boundaries of balls nor on the half-spaces of P-mass h. We also allow perturbations of measures in $\mathcal{P}^{K,h}(\mathbb{R}_d)$. A sub-Gaussian measure Q with variance $V^2 > 0$ is a measure $Q \in \mathcal{P}(\mathbb{R}_d)$ such that $Q(\mathrm{B}(0,t)^c) \leq \exp(-\frac{t^2}{2V^2})$ for all $t > V$. The set of such measures is denoted by $\mathcal{P}^{(V)}(\mathbb{R}_d)$. As well we can define $\mathcal{P}^{(V),h}(\mathbb{R}_d)$. The set $\mathcal{P}^{(V),h}(\mathbb{R}_d)$ might be thought of as perturbations of $\mathcal{P}^{K,h}(\mathbb{R}_d)$. Indeed, if $X = Y + Z$, where X has distribution in $\mathcal{P}^{K,h}(\mathbb{R}_d)$ and Z is Gaussian with variance σ^2, then Z has distribution in $\mathcal{P}^{(V),h}(\mathbb{R}_d)$, with $V = K + \sigma$. All these sets of distributions are included in $\mathcal{P}_2(\mathbb{R}_d)$, that denotes the set of distributions with finite second moment.

For all $P \in \mathcal{P}(\mathbb{R}_d)$ and $n \in \mathbb{N}^*$, $\mathbb{X}_n = \{X_1, X_2, \ldots, X_n\}$ denotes a n-sample from P, meaning that the X_i's are independent and sampled according to P. Also, $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ denotes the empirical measure associated to P, where $\delta_x \in \mathcal{P}(\mathbb{R}_d)$ is such that $\delta_x(\{x\}) = 1$. Then $\mathcal{P}_n(\mathbb{R}_d)$ is the set of $P \in \mathcal{P}(\mathbb{R}_d)$ uniform on a set of $n \in \mathbb{N}^*$ points.

An alternative definition to Definition 1, for the distance to measure, might be stated in terms of sub-measures. Let $x \in \mathbb{R}^d$. We define $\mathcal{P}_{x,h}(P)$ as the set of distributions $P_{x,h} = \frac{1}{h}Q$, for Q a sub-measure of P coinciding with P on $\mathrm{B}(x, \delta_{P,h}(x))$, and such that $Q(\mathbb{R}_d) = h$ and $\mathrm{Supp}(Q) \subset \overline{\mathrm{B}}(x, \delta_{P,h}(x))$. Note that when $P \in \mathcal{P}^{K,h}(\mathbb{R}_d)$, $\mathcal{P}_{x,h}(P)$ is reduced to a singleton $\{P_{x,h}\}$ with $P_{x,h}$ defined for all Borel sets B by $P_{x,h}(B) = \frac{1}{h}P\left(B \cap \mathrm{B}(x, \delta_{P,h}(x))\right)$. From [7, Proposition 3.3], it holds, for any $x \in \mathbb{R}^d$ and $P_{x,h} \in \mathcal{P}_{x,h}(P)$,

$$\mathrm{d}^2_{P,h}(x) = P_{x,h}\|x - u\|^2 = \|x - m(P_{x,h})\|^2 + v(P_{x,h}), \tag{1}$$

with $m(P_{x,h}) = P_{x,h}u$ the mean of $P_{x,h}$ and $v(P_{x,h}) = P_{x,h}\|u - m(P_{x,h})\|^2$ its variance. For convenience, we denote by $M(P_{x,h}) = P_{x,h}\|u\|^2$ the second moment of $P_{x,h}$, so that $M(P_{x,h}) = \|m(P_{x,h})\|^2 + v(P_{x,h})$. Whenever P is in $\mathcal{P}^{(V)}(\mathbb{R}_d)$, M satisfies the following property.

Lemma 2.
Let $P \in \mathcal{P}^{(V)}(\mathbb{R}_d)$, then $\forall\ x \in \overline{\mathbb{R}_d}$ and $h \in (0,1]$, $M(P_{x,h}) \leq \frac{2V^2}{h}$.

The proof of Lemma 2 is deferred to Section A.1.

2.2 From balls to half-spaces: structure of the local means set

In the previous part, we have seen that the DTM $\mathrm{d}_{P,h}$ is built from sub-measures of P supported on balls of P-mass h. Now, by making the center of a ball go to ∞ along a direction $v \in \mathrm{S}(0,1)$ such that the ball keeps a fixed mass h, we obtain a sub-measure of P supported on a half-space, as follows.

For $v \in \mathrm{S}(0,1)$, we denote by v_∞ the infinite point associated to the direction v. It can be seen as a limit point $\lim_{\lambda\to+\infty} \lambda v$. Then, we denote $\overline{\mathbb{R}}_d = \mathbb{R}_d \bigcup \{v_\infty \mid v \in \mathrm{S}(0,1)\}$. Note that we can equip $\overline{\mathbb{R}}_d$ with the metric $\mathrm{d}_{\overline{\mathbb{R}}_d}$ defined by $\mathrm{d}_{\overline{\mathbb{R}}_d}(x,y) = \|\phi(x) - \phi(y)\|$, with $\phi(x) = \frac{x}{\sqrt{1+\|x\|^2}}$ when $x \in \mathbb{R}_d$ and $\phi(v_\infty) = v$ for all $v \in \mathrm{S}(0,1)$. Also, for this metric, a sequence $(x_n)_{n\in\mathbb{N}}$ of $\overline{\mathbb{R}}_d$ converges to v_∞ if and only if $\lim_{n\to+\infty}\|x_n\| = +\infty$ and $\lim_{n\to+\infty}\frac{x_n}{\|x_n\|} = v$ with the convention $\frac{w_\infty}{\|w_\infty\|} = w$ for all $w \in \mathrm{S}(0,1)$.

Let $v \in \mathrm{S}(0,1)$, set $c_{P,h}(v) = \sup\{c \in \mathbb{R} \mid P\left(\{x \in \mathbb{R}_d \mid \langle x, v\rangle > c\}\right) > h\}$. Then, $\mathrm{H}(v, c_{P,h}(v))$ corresponds to the largest (for the inclusion order) half-space directed by v with P-mass at most h, which contains all the λv's for λ large enough.

Lemma 3.
Let $v \in \mathrm{S}(0,1)$ and $P \in \mathcal{P}(\mathbb{R}_d)$. Assume that $P(\partial \mathrm{H}(v, c_{P,h}(v))) = 0$. If $x_n = nv$ for all $n \in \mathbb{N}$, then for P-almost all $y \in \mathbb{R}_d$, we have:

$$\lim_{n \to +\infty} \mathbb{1}_{\mathrm{B}(x_n, \delta_{P,h}(x_n))}(y) = \mathbb{1}_{\mathrm{H}(v, c_{P,h}(v))}(y).$$

If $(x_n)_{n \in \mathbb{N}}$ is a sequence of $\mathbb{R}_d$ such that $\lim_{n \to +\infty} \mathrm{d}_{\overline{\mathbb{R}}_d}(x_n, v_\infty) = 0$, then, the result holds up to a subsequence.

The proof of Lemma 3 is given in the Appendix, Section A.2. For all $P \in \mathcal{P}_2(\mathbb{R}_d)$, we can generalize the definition of $\mathcal{P}_{x,h}(P)$, $P_{x,h}$, $m(P_{x,h})$, $v(P_{x,h})$ and $M(P_{x,h})$ to the elements $x = v_\infty \in \overline{\mathbb{R}}_d \backslash \mathbb{R}_d$ for all $v \in \mathrm{S}(0,1)$. Note that when $P \in \mathcal{P}^{K,h}(\mathbb{R}_d)$, $\mathcal{P}_{v_\infty,h}(P)$ is reduced to the singleton $\{P_{v_\infty,h}\}$ with $P_{v_\infty,h}$ equal to $\frac{1}{h} P(B \cap \mathrm{H}(v, c_{P,h}(v)))$ for all Borel set B. Intuitively, he distributions $P_{v_\infty,h}$ behave like extreme points of $\{P_{x,h} \mid x \in \mathbb{R}_d\}$. This intuition is formalized by the following Lemma. Denote $\mathcal{M}_h(P) = \left\{m(P_{x,h}) \mid x \in \overline{\mathbb{R}}_d\right\}$.

Lemma 4.
Let $P \in \mathcal{P}_2(\mathbb{R}_d)$, the set $\mathrm{Conv}(\mathcal{M}_h(P))$ is equal to $\bigcap_{v \in \mathrm{S}(0,1)} H^c(v, \langle m(P_{v_\infty,h}), v \rangle)$.

A straightforward consequence of Lemma 4 is the following Lemma 5.

Lemma 5.
Let $P \in \mathcal{P}_2(\mathbb{R}_d)$, then

$$\forall\, 0 < h < h' \leq 1,\ \mathrm{Conv}\left(\mathcal{M}_{h'}(P)\right) \subset \mathrm{Conv}\left(\mathcal{M}_h(P)\right).$$

The proofs of Lemmas 4 and 5 are to be found in Section A.3 and A.4. A key property of the local means sets $\mathcal{M}_h(P)$ is convexity. This will be of particular interest in Section 3.1. We begin with the finite-sample case.

Lemma 6.
Let $P_n \in \mathcal{P}_n(\mathbb{R}_d)$ such that $\mathrm{Supp}(P_n)$ is a set of n points in general position, as described in [3, Section 3.1.4], meaning that any subset of $\mathrm{Supp}(P_n)$ with size at most $d+1$ is a set of affinely independent points, set $q \in [\![1, n]\!]$. Then, the set $\mathcal{M}_{\frac{q}{n}}(P_n)$ is convex.

Proof
[Proof of lemma 6] Let

$$\hat{\mathcal{M}}_h(P_n) = \left\{ \bar{x} = \frac{1}{q} \sum_{p \in \mathrm{NN}_{q,\mathbb{X}_n}(x)} p \mid x \in \overline{\mathbb{R}}_d,\ \mathrm{NN}_{q,\mathbb{X}_n}(x) \in \mathcal{NN}_{q,\mathbb{X}_n}(x) \right\},$$

with $\mathcal{NN}_{q,\mathbb{X}_n}(x)$ the collection of all sets of q-nearest neighbors associated to x. Note that different $\bar{x}$ may be associated to the same x, and also note that $\hat{\mathcal{M}}_h(P_n) \subset \mathcal{M}_h(P_n)$. Moreover, $\mathrm{Conv}(\mathcal{M}_h(P_n)) = \mathrm{Conv}(\hat{\mathcal{M}}_h(P_n))$ since any $m(P_{n\,x,h})$, for $P_{n\,x,h} \in \mathcal{P}_{x,h}(P_n)$, can be expressed as a convex combination of the $\bar{x}$'s.

Then, $\mathbb{R}_d$ breaks down into a finite number of weighted Voronoï cells $\mathcal{C}_{P_n,h}(\bar{x}) = \{z \in \mathbb{R}_d \mid \|z - \bar{x}\|^2 + \hat{\omega}^2(\bar{x}) \leq \|z - \bar{y}\|^2 + \hat{\omega}^2(\bar{y}), \forall \bar{y} \in \hat{\mathcal{M}}_h(P_n)\}$, with $\hat{\omega}^2(\bar{x}) = \frac{1}{q} \sum_{p \in \mathrm{NN}_{q,\mathbb{X}_n}(x)} \|p - \bar{x}\|^2$ the weight associated to any point $\bar{x} = \frac{1}{q} \sum_{p \in \mathrm{NN}_{q,\mathbb{X}_n}(x)} p$ in $\hat{\mathcal{M}}_h(P_n)$. According to [3, Theorem 4.3], the weighted Delaunay triangulation partitions the convex hull of any finite set of weighted points $\mathbb{X}$ in general position by d-dimensional simplices with vertices in $\mathbb{X}$, provided that the associated weighted Voronoï cells of all the points in $\mathbb{X}$ are non empty. By duality, (also see [3, Lemma 4.5]) these vertices are associated to weighted Voronoï cells that have non-empty common intersection. Thus,

any $\theta \in \mathrm{Conv}(\mathcal{M}_h(P_n))$ satisfies $\theta = \sum_{i=0}^d \lambda_i \bar{x}^i$ for some $\bar{x}^i$'s in $\hat{\mathcal{M}}_h(P_n)$ and some non negative λ_i's such that $\sum_{i=0}^d \lambda_i = 1$. Also, there exists some x^* in the intersection of the $d+1$ weighted Voronoï cells, $\left(\mathcal{C}_{P_n,h}(\bar{x}^i)\right)_{i\in[\![0,d]\!]}$.

Set $P_{n\,x^*,h} := \sum_{i=0}^d \lambda_i P_i$, with $P_i = \frac{1}{q}\sum_{p\in \mathrm{NN}^i_{q,\mathbb{X}_n}(x^*)} \delta_{\{p\}}$ when $\bar{x}^i = \frac{1}{q}\sum_{p\in \mathrm{NN}^i_{q,\mathbb{X}_n}(x^*)} p$. Then, $P_{n\,x^*,h}$ is a probability measure such that $hP_{n\,x^*,h}$ ($h = \frac{q}{n}$) coincides with P_n on $\mathrm{B}(x, \delta_{P_n,h}(x))$ and is supported on $\overline{\mathrm{B}}(x, \delta_{P_n,h}(x))$. Thus it belongs to $\mathcal{P}_{x^*,h}(P_n)$. Moreover, its mean $m(P_{n\,x^*,h}) = \theta$. Thus, $\theta \in \mathcal{M}_h(P_n)$. ■

If $P \in \mathcal{P}^K(\mathbb{R}_d)$, convexity of $\mathcal{M}_h(P)$ might be deduced from the above Lemma 6 using the convergence of the empirical distribution P_n towards P in a probabilistic sense. This is summarized by the following Lemma.

Lemma 7.
Let $P \in \mathcal{P}^K(\mathbb{R}_d)$ and $\theta \in \mathrm{Conv}(\mathcal{M}_h(P))$. There exists sequences $q_n \in \mathbb{N}$, $\alpha_n \to 0$, $P_n \in \mathcal{P}_n(\mathbb{R}_d)$ with the points in $\mathrm{Supp}(P_n)$ in general position, and $y_n \in \mathrm{Conv}(\mathcal{M}_{\frac{q_n}{n}}(P_n))$ such that

i) $\frac{q_n}{n} \to h$,

ii) $\|y_n - \theta\| \le \alpha_n$,

iii) $\sup_{x\in\overline{\mathbb{R}}_d} \|m(P_{nx,\frac{q_n}{n}}) - m(P_{x,h})\| \le \alpha_n$.

Lemma 7 follows from probabilistic arguments when $\mathbb{X}_n$ is sampled at random. Its proof can be found in Section A.5. Equipped with Lemma 7, we can prove the convexity of $\mathcal{M}_h(P)$.

Proposition 8.
If $P \in \mathcal{P}^K(\mathbb{R}_d)$ for $K > 0$ is such that $P(\partial \mathrm{H}(v,c)) = 0$ and $P(\partial \mathrm{B}(x,r)) = 0$ for all $v \in \mathrm{S}(0,1)$, $x \in \mathbb{R}_d$, $c \in \mathbb{R}$, $r \ge 0$, then for all $h \in (0,1]$, $\mathcal{M}_h(P)$ is convex.

Proof
[Proof of Proposition 8] Let $\theta \in \mathrm{Conv}(\mathcal{M}_h(P))$, P_n, q_n, α_n, y_n as in Lemma 7 and for short let $h_n = \frac{q_n}{n}$.

Since $\mathcal{M}(P_n, h_n)$ is convex, there is a sequence $(x_n)_{n\ge N}$ in $\overline{\mathbb{R}}_d$ such that $y_n = m(P_{n\,x_n,h_n})$ converges to θ. If $(x_n)_{n\ge N}$ is bounded, then up to a subsequence we have $x_n \to x$, for some $x \in \mathbb{R}^d$. If not, considering $\frac{x_n}{\|x_n\|}$, up to a subsequence we get $x_n \to v_\infty$. In any case $x_n \to x$, for $x \in \overline{\mathbb{R}}_d$. Combining Lemma 7 and Lemma 3 yields $\theta = m(P_{x,h})$. Thus, $\theta \in \mathcal{M}_h(P)$. ■

2.3 The DTM defined as a power distance

A **power distance** indexed on a set I is the square root of a power function $f_{\tau,\omega}$ defined on $\mathbb{R}_d$ from a family of centers $\tau = (\tau_i)_{i\in I}$ and weights $\omega = (\omega_i)_{i\in I}$ by $f_{\tau,\omega} : x \mapsto \inf_{i\in I} \|x - \tau_i\|^2 + {\omega_i}^2$. A **$k$-power distance** is a power distance indexed on a finite set of cardinal $|I| = k$.

In [7, Proposition 3.3], the authors point out that $P_{x,h}\|x-u\|^2 \le Q\|x-u\|^2$ for all $Q \in \mathcal{P}(\mathbb{R}_d)$ such that hQ is a sub-measure of P. This remark, together with (1), provides an expression for the DTM as a power distance.

Proposition 9 ([7, Proposition 3.3])**.**
If $P \in \mathcal{P}_2(\mathbb{R}_d)$, then for all $x \in \mathbb{R}_d$, we have:

$$\mathrm{d}^2_{P,h}(x) = \inf_{y\in\overline{\mathbb{R}}_d} \inf_{P_{y,h}\in\mathcal{P}_{y,h}(P)} \|x - m(P_{y,h})\|^2 + v(P_{y,h}),$$

and the infimum is attained at $y = x$ and any measure $P_{x,h} \in \mathcal{P}_{x,h}(P)$.

As noted in Mérigot et al [10], this expression holds for the empirical DTM $\mathrm{d}_{P_n,h}$. In this case, $m(P_{n,x,h})$ corresponds to the barycentre of the $q = nh$ nearest-neighbors of x in $\mathbb{X}_n$, $\mathrm{NN}_{q,\mathbb{X}_n}(x)$, and $v(P_{n,y,h}) = \frac{1}{q}\sum_{p\in\mathrm{NN}_{q,\mathbb{X}_n}(x)} \|x-p\|^2$, at least for points x whose set of q nearest neighbors is uniquely defined.

2.4 Semiconcavity and DTM

In the following, we will often use the following lemma connected to the property of concavity of the function $x \mapsto \mathrm{d}^2_{P,h}(x) - \|x\|^2$.

Lemma 10 ([7, Proposition 3.6])**.**
If $P \in \mathcal{P}(\mathbb{R}_d)$, then for all x, $y \in \mathbb{R}_d$ and $P_{x,h} \in \mathcal{P}_{x,h}(P)$,

$$\mathrm{d}^2_{P,h}(y) - \|y\|^2 \leq \mathrm{d}^2_{P,h}(x) - \|x\|^2 - 2\langle y-x, m(P_{x,h})\rangle,$$

with equality if and only if $P_{x,h} \in \mathcal{P}_{y,h}(P)$.

3 The k-PDTM: a coreset for the DTM

In Proposition 9, we have written the DTM as a power distance. This remark has already been exploited in [10] and [6], where the DTM has been approximated by n-power distances. In this paper, we propose to keep only k centers.

Definition 11.
For any $P \in \mathcal{P}_2(\mathbb{R}_d)$, we define $\mathcal{O}pt(P,h,k)$ by:

$$\mathcal{O}pt(P,h,k) = \arg\min\left\{ P \min_{i\in[\![1,k]\!]} \|u - m(P_{t_i,h})\|^2 + v(P_{t_i,h}) \mid t = (t_1,t_2,\dots t_k) \in \overline{\mathbb{R}}_d^{(k)} \right\}.$$

A closely related notion to Definition 11 is the following weighted Voronoï measures.

Definition 12.
*A set of **weighted Voronoï measures** associated to a distribution $P \in \mathcal{P}_2(\mathbb{R}_d)$, $t \in \overline{\mathbb{R}}_d^{(k)}$ and $h \in (0,1]$ is a set $\left\{\tilde{P}_{t_1,h}, \tilde{P}_{t_2,h}, \dots \tilde{P}_{t_k,h}\right\}$ of $k \in \mathbb{N}^*$ positive sub-measures of P such that $\sum_{i=1}^k \tilde{P}_{t_i} = P$ and*

$$\forall x \in \mathrm{Supp}(\tilde{P}_{t_i,h}),\ \|x - m(P_{t_i,h})\|^2 + v(P_{t_i,h}) \leq \|x - m(P_{t_j,h})\|^2 + v(P_{t_j,h}),\ \forall j \in [\![1,k]\!].$$

We denote by $\tilde{m}(\tilde{P}_{t_i,h}) = \frac{\tilde{P}_{t_i,h}u}{\tilde{P}_{t_i,h}(\mathbb{R}_d)}$ the expectation of $\tilde{P}_{t_i,h}$, with the convention $\tilde{m}(\tilde{P}_{t_i,h}) = 0$ when $\tilde{P}_{t_i,h}(\mathbb{R}_d) = 0$.

Note that a set of weighted Voronoï measures can always be assigned to any $P \in \mathcal{P}_2(\mathbb{R}_d)$ and $t \in \overline{\mathbb{R}}_d^{(k)}$, it suffices to split $\mathbb{R}_d$ in weighted Voronoï cells associated to the centers $(m(P_{t_i,h}))_{i\in[\![1,k]\!]}$ and weights $(v(P_{t_i,h}))_{i\in[\![1,k]\!]}$, see [3, Section 4.4.2], and split the remaining mass on the border of the cells in a measurable arbitrary way.

Theorem 13.
For all $h \in (0,1]$, $k \in \mathbb{N}^*$ and $P \in \mathcal{P}^K(\mathbb{R}_d)$ for some $K > 0$, such that $P(\partial \mathrm{H}(v, c_{P,h}(v))) = 0$ for all $v \in \mathrm{S}(0,1)$, the set $\mathcal{O}pt(P,h,k)$ is not empty. Moreover, there is some $s \in \mathcal{O}pt(P,h,k) \cap \overline{\mathrm{B}}(0,K)^{(k)}$ such that $s_i = \tilde{m}(\tilde{P}_{s_i,h})$ for all $i \in [\![1,k]\!]$.

Proof
[Sketch of proof] For $s \in \overline{\mathbb{R}}_d^{(k)}$, set $f_s : x \in \mathbb{R}_d \mapsto \min_{i\in[\![1,k]\!]} \left(\|x - m(P_{s_i,h})\|^2 + v(P_{s_i,h})\right)$. Then, Lemma 3 and the dominated convergence theorem yield $\inf_{t\in\mathbb{R}_d} Pf_t(u) = \inf_{t\in\overline{\mathbb{R}}_d} Pf_t(u)$.

Let $(t_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}_d^{(k)}$ such that $Pf_{t_n}(u) \leq \inf_{t\in\overline{\mathbb{R}}_d} Pf_t(u) + \frac{1}{n}$, and denote by m^* the limit of a converging subsequence of $\left(\tilde{m}(\tilde{P}_{t_{n,1},h}), \tilde{m}(\tilde{P}_{t_{n,2},h}), \ldots, \tilde{m}(\tilde{P}_{t_{n,k},h})\right)_{n\in\mathbb{N}}$ in the compact space $\overline{\mathrm{B}}(0,K)^{(k)}$. Then, Lemma 10 and (1) yield $Pf_{m^*}(u) = \inf_{t\in\overline{\mathbb{R}}_d} Pf_t(u)$.

Set $s_i = \tilde{m}(\tilde{P}_{m_i^*,h})$ for all $i \in [\![1,k]\!]$, then $\tilde{m}(\tilde{P}_{s_i,h}) = s_i$ and $Pf_{m^*}(u) = Pf_s(u)$. ■

The detailed proof of Theorem 13 is given in Section B.1. Note that the distributions in $\mathcal{P}^{K,h}$ are in the scope of Theorem 13.

3.1 Two equivalent definitions for the k-PDTM

Definition 14.
*Let $P \in \mathcal{P}_2(\mathbb{R}_d)$, the k-**power distance to a measure** (k-PDTM) $\mathrm{d}_{P,h,k}$ is defined for any $s \in \mathcal{O}pt(P,h,k)$ by:*

$$\mathrm{d}^2_{P,h,k}(x) = \min_{i\in[\![1,k]\!]} \|x - m(P_{s_i,h})\|^2 + v(P_{s_i,h}).$$

*An ϵ-**approximation of the k-PDTM**, denoted by an $\mathrm{d}^2_{P,h,k,\epsilon}$ is a function defined by the previous expression but for some $s \in \overline{\mathbb{R}}_d^{(k)}$ satisfying*

$$P \min_{i\in[\![1,k]\!]} \|u - m(P_{s_i,h})\|^2 + v(P_{s_i,h}) \leq \inf_{t\in\overline{\mathbb{R}}_d^{(k)}} P \min_{i\in[\![1,k]\!]} \|u - m(P_{t_i,h})\|^2 + v(P_{t_i,h}) + \epsilon.$$

Theorem 13 states that the k-PDTM is well defined when $P \in \mathcal{P}^K(\mathbb{R}_d)$ and satisfies $P(\partial \mathrm{H}_{P,h}(v, c_{P,h}(v))) = 0$ for all $v \in \mathrm{S}(0,1)$. Nonetheless, whenever $\mathcal{O}pt(P,h,k)$ is not a singleton, the k-PDTM is not unique. Note that for all $x \in \mathbb{R}_d$, $\mathrm{d}_{P,h,k}(x) \geq \mathrm{d}_{P,h}(x)$.

Definition 15.
The set $\mathrm{OPT}(P,h,k)$ is defined by:

$$\mathrm{OPT}(P,h,k) = \arg\min \left\{ P \min_{i\in[\![1,k]\!]} \|u - \tau_i\|^2 + \omega^2_{P,h}(\tau_i) \mid \tau = (\tau_1, \tau_2, \ldots \tau_k) \in \mathbb{R}_d^{(k)} \right\},$$

with $\omega_{P,h}(\tau) = \inf\left\{\omega > 0 \mid \forall x \in \mathbb{R}_d,\ \|x-\tau\|^2 + \omega^2 \geq \mathrm{d}^2_{P,h}(x)\right\}$ for $\tau \in \mathbb{R}_d$, that is:

$$\omega^2_{P,h}(\tau) = \sup_{x\in\mathbb{R}_d} \mathrm{d}^2_{P,h}(x) - \|x-\tau\|^2. \tag{2}$$

The following Lemma shows that $\mathrm{OPT}(P,h,k)$ is included in $\mathrm{Conv}(\mathcal{M}_h(P))^{(k)}$.

Lemma 16.
Let $P \in \mathcal{P}^K(\mathbb{R}_d) \cup \mathcal{P}^{(V)}(\mathbb{R}_d)$. Then $\theta \in \mathrm{Conv}(\mathcal{M}_h(P))$ if and only if $\omega_{P,h}(\theta) < +\infty$.

Proof
[Proof of Lemma 16] According to Proposition 9, for all $x \in \mathbb{R}_d$, $\mathrm{d}^2_{P,h}(x) - \|x-\theta\|^2$ may be written as

$$\inf_{y\in\overline{\mathbb{R}}_d} \inf_{P_{y,h}\in\mathcal{P}_{y,h}(P)} \left\{\|m(P_{y,h})\|^2 + v(P_{y,h}) - \|\theta\|^2 + 2\langle x, \theta - m(P_{y,h})\rangle\right\},$$

which is lower-bounded by

$$\inf_{y\in\overline{\mathbb{R}}_d} \inf_{P_{y,h}\in\mathcal{P}_{y,h}(P)} \left\{\|m(P_{y,h})\|^2 + v(P_{y,h})\right\} - \|\theta\|^2 + \inf_{\tau\in\mathcal{M}_h(P)} \left\{2\langle x, \theta - \tau\rangle\right\}.$$

Assume $\theta \notin \mathrm{Conv}(\mathcal{M}_h(P))$. According to Lemma 4, $\mathrm{Conv}(\mathcal{M}_h(P))$ is a convex and compact subset of $\mathbb{R}_d$. The Hahn-Banach separation theorem thus provides some vector $v \in \mathbb{R}_d$ and $C > 0$ such that $\forall \tau \in \mathcal{M}_h(P)$, $\langle \theta - \tau, v \rangle < C$. Setting $x_n = -nv$ for $n \in \mathbb{N}^*$ yields $\lim_{n\to+\infty} \inf_{\tau\in\mathcal{M}_h(P)} \langle x_n, \theta - \tau\rangle = +\infty$. Thus, $\sup_{x\in\mathbb{R}_d} \mathrm{d}^2_{P,h}(x) - \|x-\theta\|^2 = +\infty$.

Now, let $\theta \in \mathrm{Conv}(\mathcal{M}_h(P))$, we can write $\theta = \sum_{i=0}^d \lambda_i m(P_i)$ for $P_i = P_{x_i,h}$ with the x_i's in $\overline{\mathbb{R}}_d$. We have:

$$\begin{aligned}
\sup_{x\in\overline{\mathbb{R}}_d} \mathrm{d}^2_{P,h}(x) - \|x-\theta\|^2 &= \sup_{x\in\overline{\mathbb{R}}_d} \sum_{i=0}^d \lambda_i (\mathrm{d}^2_{P,h}(x) - \|x-\theta\|^2) \\
&\leq \sup_{x\in\overline{\mathbb{R}}_d} \sum_{i=0}^d \lambda_i \left(\|x-m(P_i)\|^2 + v(P_i) - \|x-\theta\|^2\right) \\
&= \sup_{x\in\overline{\mathbb{R}}_d} \sum_{i=0}^d \lambda_i \left(v(P_i) + 2\langle x, \theta - m(P_i)\rangle + \|m(P_i)\|^2 - \|\theta\|^2\right) \\
&= \sum_{i=0}^d \lambda_i \left(v(P_i) + \|m(P_i)\|^2 - \|\theta\|^2\right),
\end{aligned}$$

according to Proposition 9. Thus, we get that

$$\omega^2_{P,h}(\theta) + \|\theta\|^2 \leq \sum_{i=0}^d \lambda_i (v(P_i) + \|m(P_i)\|^2) \leq \sup_{x\in\overline{\mathbb{R}}_d} \left\{v(P_{x,h}) + \|m(P_{x,h})\|^2\right\}. \qquad (3)$$

Lemma 2 yields $\omega^2_{P,h}(\theta) < +\infty$. ■

Theorem 17.
If $P \in \mathcal{P}^{K,h}(\mathbb{R}_d)$ for some $h \in (0,1]$ and $K > 0$, or $P \in \mathcal{P}_n(\mathbb{R}_d)$ such that $\mathrm{Supp}(P_n)$ is a set of n points in general position as described in Lemma 6, for some $h = \frac{q}{n}$ with $q \in [\![1, n]\!]$, then, any function $\mathrm{d}_{P,h,k}$ satisfies for some $\theta \in \mathrm{OPT}(P,h,k)$:

$$\mathrm{d}^2_{P,h,k}(x) = \min_{i\in[\![1,k]\!]} \|x-\theta_i\|^2 + \omega^2_{P,h}(\theta_i),\ \forall x \in \mathbb{R}_d.$$

Conversely, for all $\theta \in \mathrm{OPT}(P,h,k)$, $x \mapsto \sqrt{\min_{i\in[\![1,k]\!]} \|x-\theta_i\|^2 + \omega^2_{P,h}(\theta_i)}$ is a k-PDTM.

Proof
[Proof of Theorem 17] For all $\tau \in \mathbb{R}_d^{(k)}$, for all $i \in [\![1,k]\!]$, if $\tau_i \notin \mathrm{Conv}(\mathcal{M}_h(P))$, then according to Lemma 16, $\omega_{P,h}(\tau_i) = +\infty$. In this case, $\tau \notin \mathrm{OPT}(P,h,k)$.
Thus, for all $\tau \in \mathrm{OPT}(P,h,k)$, for all i, $\tau_i \in \mathrm{Conv}(\mathcal{M}_h(P))$. According to Proposition 8 and Lemma 6, $\mathcal{M}_h(P)$ is convex. Thus,

$$\mathrm{OPT}(P,h,k) = \arg\min \left\{ P \min_{i\in[\![1,k]\!]} \|u-\tau_i\|^2 + \omega^2_{P,h}(\tau_i) \mid \tau = (\tau_1, \tau_2, \dots \tau_k) \in \mathcal{M}_h(P)^{(k)} \right\}.$$

Moreover, according to Proposition 9, and (2), $\omega^2_{P,h}(m(P_{t,h})) = v(P_{t,h})$, for all $t \in \overline{\mathbb{R}}_d$. Thus,

$$\inf_{t\in\overline{\mathbb{R}}_d^{(k)}} P \min_{i\in[\![1,k]\!]} \|x-m(P_{t_i,h})\|^2 + v(P_{t_i,h}) = \inf_{\tau\in\mathbb{R}_d^{(k)}} P \min_{i\in[\![1,k]\!]} \|u-\tau_i\|^2 + \omega^2_{P,h}(\tau_i).$$

■

Therefore, Theorem 17 allows to consider the function $\mathrm{d}_{P,h,k}$ as the square root of a minimizer of the $L_1(P)$ norm $f \mapsto P|f - \mathrm{d}^2_{P,h}|(u)$ among all the k-power functions f which graph lies above the graph of the function $\mathrm{d}^2_{P,h}$.

3.2 Proximity to the DTM

Here we show that the k-PDTM approximates the DTM in the following sense.

Proposition 18.
Let $P \in \mathcal{P}^K(\mathbb{R}_d)$ for $K > 0$ and let $M \subset \mathrm{B}(0, K)$ be such that $P(M) = 1$. Let $f_M(\varepsilon)$ denote the ε covering number of M. Then we have

$$0 \leq P\mathrm{d}^2_{P,h,k}(u) - \mathrm{d}^2_{P,h}(u) \leq 2f_M^{-1}(k)\zeta_{P,h}(f_M^{-1}(k)), \quad \text{with} \quad f_M^{-1}(k) = \inf\{\varepsilon > 0 | \quad f_M(\varepsilon) \leq k\},$$

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

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

Proof
[Proof of Proposition 18] The first inequality comes from Proposition 9.

We then focus on the second bound. By definition of $\mathrm{d}_{P,h,k}$, for all $x \in \mathbb{R}_d$ and $t = (t_1, t_2, \ldots, t_k) \in \mathbb{R}_d^{(k)}$ we have: $P\mathrm{d}^2_{P,h,k}(x) \leq P \min_{i \in [\![1,k]\!]} \|u - m(P_{t_i,h})\|^2 + v(P_{t_i,h})$. Thus,

$$\begin{aligned}
P\mathrm{d}^2_{P,h,k}(u) - \mathrm{d}^2_{P,h}(u) &\leq P \min_{i \in [\![1,k]\!]} \|u - m(P_{t_i,h})\|^2 + v(P_{t_i,h}) - \mathrm{d}^2_{P,h}(u) \\
&= P \min_{i \in [\![1,k]\!]} (\mathrm{d}^2_{P,h}(t_i) - \|t_i\|^2) - (\mathrm{d}^2_{P,h}(u) - \|u\|^2) + \langle u - t_i, -2m(P_{t_i,h})\rangle \\
&\leq P \min_{i \in [\![1,k]\!]} 2\langle u - t_i, m(P_{u,h}) - m(P_{t_i,h})\rangle \\
&\leq 2P \min_{i \in [\![1,k]\!]} \|u - t_i\| \|m(P_{u,h}) - m(P_{t_i,h})\|,
\end{aligned}$$

where we used (1), Lemma 10 and Cauchy-Schwarz inequality. Now choose $t_1, \ldots, t_k$ as a $f_M^{-1}(k)$-covering of M. The result follows. ■

When 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 focus on two cases in which Proposition 18 may be adapted. First, the case where the distribution P has an ambient-dimensional support is investigated.

Corollary 19.
Assume that P have a density f satisfying $0 < f_{min} \leq f \leq f_{max}$. Then

$$0 \leq P\mathrm{d}^2_{P,h,k}(u) - \mathrm{d}^2_{P,h}(u) \leq C_{f_{max},K,d,h} k^{-2/d}.$$

The proof of Corollary 19 is given in Section B.2. Note that no assumptions on the geometric regularity of M is required for Corollary 19 to hold. In the case where M has a lower-dimensional structure, more regularity is required, as stated by the following corollary.

Corollary 20.
Suppose that P is supported on a compact d'-dimensional $\mathcal{C}^2$-submanifold of $\mathrm{B}(0, K)$, denoted by N. Assume that N has positive reach ρ, and 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,

$$P(\mathrm{B}(x, r)) \geq c f_{min} r^{d'} \wedge 1. \tag{4}$$

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

Note that (4), also known as (cf_{min}, d')-standard assumption, is usual in set estimation (see, e.g., [8]). 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., [2, Corollary 1]). An important feature of Corollary 20 is that this approximation bound does not depend on the ambient dimension. The proof of Corollary 20 may be found in Section B.3.

3.3 Wasserstein stability for the k-PDTM

Next we assess that our k-PDTM shares with the DTM the key property of robustness to noise.

Proposition 21.
Let $P \in \mathcal{P}^K(\mathbb{R}_d)$ for some $K > 0$, $Q \in \mathcal{P}_2(\mathbb{R}_d)$, and $\epsilon > 0$. Set $\mathrm{d}^2_{Q,h,k,\epsilon}$ an ϵ-approximation of the k-PDTM of Q, then $P\left|\mathrm{d}^2_{Q,h,k,\epsilon}(u) - \mathrm{d}^2_{P,h}(u)\right|$ is bounded from above by $B_{P,Q,h,k,\epsilon}$ with

$$B_{P,Q,h,k,\epsilon} = \epsilon + 3\|\mathrm{d}^2_{Q,h} - \mathrm{d}^2_{P,h}\|_{\infty,\mathrm{Supp}(P)} + P\mathrm{d}^2_{P,h,k}(u) - \mathrm{d}^2_{P,h}(u) + 2W_1(P,Q) \sup_{s\in\overline{\mathbb{R}}_d} \|m(P_{s,h})\|.$$

Note that Lemma 2 gives a bound on $m(Q_{s,h})$ whenever Q is sub-Gaussian. Also, upper-bounds for the deviation of the k-PDTM to the DTM associated to P have been derived in the previous subsection.

Proof
[Sketch of proof] For $x \in \mathrm{Supp}(P)$, $\max\left\{0, -\left(\mathrm{d}^2_{Q,h,k} - \mathrm{d}^2_{P,h}\right)(x)\right\} \leq \|\mathrm{d}^2_{P,h} - \mathrm{d}^2_{Q,h}\|_{\infty,\mathrm{Supp}(P)}$. Set $f_{Q,p}(x) = 2\langle x, m(Q_{p,h})\rangle + v(Q_{p,h})$ for $p \in \mathbb{R}_d$, and let $t \in \mathcal{O}pt(Q,h,k)$. Then, $P - Q\min_{i\in[\![1,k]\!]} f_{Q,t_i}(u) \leq 2W_1(P,Q)\sup_{t\in\overline{\mathbb{R}}_d} m(Q_{t,h})$ and for s given by Theorem 13, that is $s \in \mathcal{O}pt(P,h,k) \cap \overline{\mathrm{B}}(0,K)^{(k)}$ such that $s_i = \tilde{m}(\tilde{P}_{s_i,h})$ for all $i \in [\![1,k]\!]$, $P\min_{i\in[\![1,k]\!]} f_{Q,s_i}(u) - \min_{i\in[\![1,k]\!]} f_{P,s_i}(u)$ is bounded from above by $\|\mathrm{d}^2_{P,h} - \mathrm{d}^2_{Q,h}\|_{\infty,\mathrm{Supp}(P)}$. ■

The details of the proof of Proposition 21 can be found in Section B.4.

3.4 Geometric inference with the k-PDTM

As detailed in [7, Section 4], under suitable assumptions, the sublevel sets of the distance to measure are close enough to the sublevel sets of the distance to its support. Thus they allow to infer the geometric structure of the support. As stated below, this is also the case when replacing the distance to measure with the k-PDTM.

Proposition 22.
Let M be a compact set in $\mathrm{B}(0,K)$ such that $P(M) = 1$. Moreover, assume that there exists d' such that, for every $p \in M$ and $r \geq 0$,

$$P(\mathrm{B}(p,r)) \geq C(P) r^{d'} \wedge 1. \tag{5}$$

Let Q be a probability measure (thought of as a perturbation of P), and let Δ_P^2 denote $P\mathrm{d}^2_{Q,h,k,\varepsilon}(u)$. Then, we have

$$\sup_{x\in\mathbb{R}_d} |d_{Q,h,k,\varepsilon}(x) - d_M(x)| \leq C(P)^{-\frac{1}{d'+2}} \Delta_P^{\frac{2}{d'+2}} + W_2(P,Q) h^{-\frac{1}{2}},$$

where W_2 denotes the Wasserstein distance.

Proposition 22, whose proof can be found in Section B.5, ensures that the k-PDTM achieves roughly the same performance as the distance to measure (see, e.g., [7, Corollary 4.8]) provided that $d^2_{Q,h,k,\varepsilon}$ 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.

4 Approximation of the k-PDTM from point clouds

Let $P \in \mathcal{P}_2(\mathbb{R}_d)$, an approximation of the k-PDTM $\mathrm{d}_{P,h,k}$, is given by the **empirical k-PDTM** $\mathrm{d}_{P_n,h,k}$. Note that when $k = n$, $\mathrm{d}_{P_n,h,n}$ is equal to the q-witnessed distance. Also, when $h = 0$ we recover the k-means method.

4.1 An algorithm for the empirical k-PDTM

The following algorithm is inspired by the Lloyds algorithm. We assume that the mass parameter $h = \frac{q}{n}$ for some positive integer q. And for any $t \in \mathbb{R}_d$, we use the notation $c(t) = \frac{1}{q}\sum_{i=1}^{q} X_i(t)$, where $X_i(t)$ is one of the i-th nearest neighbor of t in $\mathbb{R}_d$. We denote $\omega^2(t) = \frac{1}{q}\sum_{i=1}^{q}\left(X_i(t) - c(t)\right)^2$, and $\mathcal{C}(t)$ the weighted Voronoï cell associated to t. We use the notation $|\mathcal{C}(t)|$ for the cardinal of $\mathcal{C}(t) \cap \mathbb{X}_n$.

Algorithm 1: Local minimum algorithm

Input : $\mathbb{X}_n$ a n-sample from P, q and k ;
`# Initialization`
Sample t_1, t_2,... t_k from $\mathbb{X}_n$ without replacement. ;
while the t_is vary make the following two steps :
  `# Decomposition in weighted Voronoi cells.`
  for j **in** $1..n$:
    Add X_j to the $\mathcal{C}(t_i)$ (for i as small as possible) satisfying $\|X_j - c(t_i)\|^2 + \omega^2(t_i) \leq \|X_j - c(t_l)\|^2 + \omega^2(t_l)\, \forall l \neq i$;
  `# Computation of the new centers and weights.`
  for i **in** $1..k$:
    $t_i = \frac{1}{|\mathcal{C}(t_i)|}\sum_{X \in \mathcal{C}(t_i)} X$;
Output : $(t_1, t_2, \ldots, t_k)$

The following proposition relies on the same arguments as in the proof of Theorem 13.

Proposition 23.
This algorithm converges to a local minimum of $t \mapsto P_n \min_{i \in [\![1,k]\!]} \|x - m(P_{n\,t_i,h})\|^2 + v(P_{n\,t_i,h})$.

The proof of Proposition 23 can be found in Section C.1.

4.2 Proximity between the k-PDTM and its empirical version

Theorem 24.
Let P be supported on $M \subset \mathrm{B}(0, K)$. Assume that we observe $X_1, \ldots, X_n$ such that $X_i = Y_i + Z_i$, where Y_i is an i.i.d n-sample 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 $1 - 7n^{-p}$, we have

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

A proof of Theorem 24 is given in Section C.2. Theorem 24, combined with Proposition 21, allows to choose k in order to minimize $P\mathrm{d}^2_{Q_{n,h,k}}(u)$. Indeed, in the framework of Corollaries 19 and 20 where the support has intrinsic dimension d', such a minimization boils down to optimize a quantity of the form

$$\frac{C\sqrt{k}K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C_P k^{-\frac{2}{d'}}.$$

Hence the choice $k \sim n^{\frac{d'}{d'+4}}$ ensures that for n large enough, only $n^{\frac{d'}{d'+4}}$ points are sufficient to approximate well the sub-level sets of the distance to support. For surface inference ($d' = 2$), this amounts to compute the distance to $n^{\frac{1}{3}}$ 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's. In some sense, Theorem 24 advocates only an upper bound on k, above which no increase of precision can be expected.

4.3 Some numerical illustration

As in [10], we sampled $n = 6000$ points from the uniform measure on a sideways with radius $\sqrt{2}$ and $\sqrt{\frac{9}{8}}$ convolved with a Gaussian $\mathcal{N}(0, \sigma^2)$ with $\sigma = 0.45$. We then plotted in grey the r-sub-level set of the q-witnessed distance and in purple, the r-sub-level set of an approximation of $\mathrm{d}_{P_n,q,k}$ with $r = 0.24$ and $q = 50$ nearest-neighbors. The approximation of $\mathrm{d}_{P_n,q,k}$ is obtained after running our algorithm 10 times and keeping the best (for the $L_1(P_n)$ loss) function obtained, each time after at most 10 iterations. Choosing $k = 100$ points leads to a too sparse approximation of the sublevel sets of the

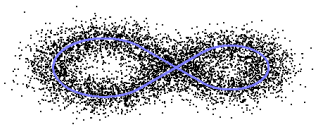

Figure 1: 6000-sample

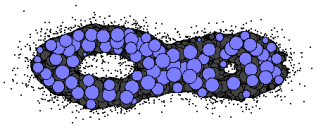

Figure 2: $k = 100$

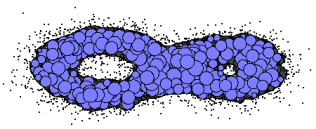

Figure 3: $k = 300$

q-witnessed distance. On the contrary, small holes which appeared in the r-sub-level set, when $k = 300$, will disappear quickly when the radius r will get larger, before the two holes get filled.

The authors are grateful to Pascal Massart, Frédéric Chazal and Marc Glisse for their precious advice.

References

[1] Eddie Aamari and Clément Levrard. "Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction". In: *ArXiv e-prints, 1512.02857* (2015).

[2] Catherine Aaron and Alejandro Cholaquidis. "On boundary detection". In: *ArXiv e-prints, 1603.08460* (2016).

[3] Jean-Daniel Boissonnat, Frédéric Chazal, and Mariette Yvinec. "Geometric and Topological Inference". In: *Cambridge University Press* (2017).

[4] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. “Theory of classification: a survey of some recent advances”. In: *ESAIM. Probability and Statistics* 9 (2005), pp. 323–375.

[5] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. 2013.

[6] Mickaël Buchet et al. “Efficient and robust persistent homology for measures”. In: *Computational Geometry* 58 (2016), pp. 70–96.

[7] Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. “Geometric Inference for Probability Measures”. In: *Foundations of Computational Mathematics* 11.6 (2011), pp. 733–751.

[8] Frédéric Chazal et al. “Convergence Rates for Persistence Diagram estimation in Topological Data Analysis”. In: *Journal of Machine Learning Research* 16 (2015), pp. 3603–3635.

[9] Herbert Edelsbrunner. *Weighted alpha-shapes*. Tech. rep. UIUCDCS-R-92-1760. Department of Computer Science, University of Illinois, 1992.

[10] Leonidas Guibas, Dmitriy Morozov, and Quentin Mérigot. “Witnessed k-distance”. In: *Discrete and Computational Geometry* 49.1 (2013), pp. 22–45.

[11] Shahar Mendelson and Roman 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: http://dx.doi.org/10.1007/s00222-002-0266-3.

[12] Quentin Mérigot. “Lower bounds for k-distance approximation”. In: *International Symposium on Computational Geometry (SoCG)*. 2013, 435—440.

[13] Jeff M. Phillips, Bei Wangy, and Yan Zheng. “Geometric Inference on Kernel Density Estimates”. In: *International Symposium on Computational Geometry (SoCG)*. 2015.

Appendix

A Proofs for Section 2

A.1 Proof of Lemma 2

Proof
Let $P \in \mathcal{P}^{(V)}(\mathbb{R}_d)$, x in $\overline{\mathbb{R}}_d$ and $P_{x,h}$ a sub-measure of P, supported on $\overline{\mathrm{B}}(x, \delta_{P,h}(x))$ (or on $\mathrm{H}(v, c_{P,h}(v))$ if $x = v_\infty \in \overline{\mathbb{R}}_d \backslash \mathbb{R}_d$), coinciding with P on $\mathrm{B}(x, \delta_{P,h}(x))$, and such that $P_{x,h}(\mathbb{R}_d) = h$. We may write

$$\begin{aligned} P_{x,h}\|u\|^2 &\leq \frac{1}{h} P\|u\|^2 \\ &\leq \frac{1}{h}\left[P\|u\|^2 \mathbb{1}_{\|u\|\leq V} + P\|u\|^2 \mathbb{1}_{\|u\|>V}\right] \\ &\leq \frac{V^2}{h} + \frac{P\|u\|^2 \mathbb{1}_{\|u\|>V}}{h}. \end{aligned}$$

Since $P\|u\|^2 \mathbb{1}_{\|u\|>V} \leq N_{V^2} t^2 \mathbb{1}_{t>V} \leq N_{V^2} t^2 = V^2$, where N_{V^2} denotes the distribution of a Gaussian distribution with variance V^2, the result of Lemma 2 follows. ■

A.2 Proof of Lemma 3

Proof
Note that for any point $x \in \partial \mathrm{H}\left(v, c_{P,h}(v)\right) \cap \mathrm{S}(0, K)$, $x = w + c_{P,h}^2(v)v$ for some w orthogonal to v. Moreover, $\|x - c_{P,h}(v)v\|^2 = K^2 - c_{P,h}^2(v)$ and $\|x - x_n\|^2 = (n - c_{P,h}(v))^2 + K^2 - c_{P,h}^2(v)$ for $x_n = nv$. Thus, we get that:

$$\mathrm{B}\left(x_n, n - c_{P,h}(v)\right) \subset \mathrm{H}\left(v, c_{P,h}(v),\right)$$

and

$$\mathrm{H}\left(v, c_{P,h}(v)\right) \cap \mathrm{Supp}(P) \subset \mathrm{B}\left(x_n, \sqrt{K^2 - c_{P,h}^2(v) + (n - c_{P,h}(v))^2}\right).$$

In particular, since $P\left(\mathrm{H}\left(v, c_{P,h}(v)\right)\right) < h$, $\mathrm{B}\left(x_n, n - c_{P,h}(v)\right) \subset \mathrm{B}(x_n, \delta_{P,h}(x))$ and since $P\left(\overline{\mathrm{H}}\left(v, c_{P,h}(v)\right)\right) \geq h$, $\mathrm{B}(x_n, \delta_{P,h}(x)) \subset \mathrm{B}\left(x_n, \sqrt{K^2 - c_{P,h}^2(v) + (n - c_{P,h}(v))^2}\right)$, with $\delta_{P,h}(x)$ the pseudo-distance defined in Section 2. Finally, for all $y \in \mathbb{R}_d$, if $\langle y, v\rangle = c_{P,h}(v) - \epsilon$ for some $\epsilon > 0$, then $\|y - x_n\|^2 = \|y\|^2 + n^2 - 2n(c_{P,h}(v) - \epsilon)$, which is superior to $K^2 + (n - c_{P,h}(v))^2 - c_{P,h}^2(v)$ for n large enough. Thus, for all n large enough, $y \notin \mathrm{B}(x_n, \delta_{P,h}(x_n))$. If $\langle y, v\rangle = c_{P,h}(v) + \epsilon$ for some $\epsilon > 0$, then $\|y - x_n\|^2 = \|y\|^2 + (n - c_{P,h}(v))^2 - c_{P,h}^2(v) - 2n\epsilon$ which is inferior to $(n - c_{P,h}(v))^2$ for n large enough. Thus, for all n large enough, $y \in \mathrm{B}(x_n, \delta_{P,h}(x_n))$, which concludes the first part of the Lemma.

Let $(x_n)_{n\geq 0}$ be a sequence in $\mathbb{R}_d$ such that $\lim_{n\to+\infty} \mathrm{d}_{\overline{\mathbb{R}}_d}(x_n, v_\infty) = 0$, that is such that $\lim_{n\to+\infty} \|x_n\| = +\infty$ and $\lim_{n\to+\infty} \frac{x_n}{\|x_n\|} = v$. Then,

$$\|x_n\| - K \leq \delta_{P,h}(x_n) \leq \|x_n\| + K.$$

Let $y \in \mathbb{R}^d$. Then,

$$\|y - x_n\|^2 - \delta_{P,h}(x_n)^2 = \|y\|^2 - 2\langle x_n, y\rangle + O(\|x_n\|) = \|x_n\|\left(\frac{\|y\|^2}{\|x_n\|} - 2\left\langle \frac{x_n}{\|x_n\|}, y\right\rangle + O(1)\right).$$

The notation $y_n = O(\|x_n\|)$ means that $\left(\frac{y_n}{\|x_n\|}\right)_{n\in\mathbb{N}}$ is bounded. Thus, up to a subsequence,

$$\lim_{n\to+\infty} \frac{\|y - x_n\|^2 - \delta_{P,h}(x_n)^2}{\|x_n\|} = 2c - 2\langle v, y\rangle,$$

for some $c \in \mathbb{R}$. We deduce that, for all $y \in \mathbb{R}_d \backslash \partial \mathrm{H}(v, c)$,

$$\mathbb{1}_{\mathrm{B}(x_n, \delta_{P,h}(x_n))}(y) \to \mathbb{1}_{\mathrm{H}(v,c)}(y).$$

In particular, $P(\mathrm{H}(v, c)) \leq h$ and $P(\overline{\mathrm{H}}(v, c)) \geq h$. Therefore, for P-almost y, $\mathbb{1}_{\mathrm{H}(v,c)}(y) = \mathbb{1}_{\mathrm{H}(v,c_{P,h}(v))}(y)$, the result then holds for $c = c_{P,h}(v)$. ■

A.3 Proof of Lemma 4

Proof

Recall that a k-extreme point x of a convex set S is a point x which lies in the interior of a k-dimensional convex set within S, but not a $k+1$-dimensional convex set within S. We will prove that the set of k-extreme points in $\mathcal{M}_h(P)$ of $\mathcal{M}_h(P)$ for $k < d$ is equal to $\{m(P_{x,h}) \mid x \in \overline{R}_d \backslash \mathbb{R}_d\}$. In particular, this will yield that the set $\mathrm{Conv}(\mathcal{M}_h(P))$ is equal to $\bigcap_{v\in \mathrm{S}(0,1)} H^c(v, \langle m(P_{v_\infty,h}), v\rangle)$.

Let $v \in \mathrm{S}(0, 1)$, then by definition the measure $P_{v_\infty,h}$ is supported on $\overline{\mathrm{H}}(v, c_{P,h}(v))$, satisfies that $hP_{v_\infty,h}$ is a sub-measure of P and the measures $hP_{v_\infty,h}$ and P coincide on $\mathrm{H}(v, c_{P,h}(v))$. Note that $P(\overline{\mathrm{H}}(v, c_{P,h}(v))) \geq h$ and $P(\mathrm{H}(v, c_{P,h}(v))) \leq h$.

We will denote $\overline{C}(P_{v_\infty,h}) = \langle m(P_{v_\infty,h}), v\rangle$, that is, $\overline{C}(P_{v_\infty,h}) = P_{v_\infty,h}\langle u, v\rangle$.

Then, for all $x \in \overline{\mathbb{R}}_d$, we decompose any measure $P_{x,h}$ as $P_1 + P_2$ with $P_1(B) = P_{x,h}(B \cap H(v, c_{P,h}(v)))$ and $P_2(B) = P_{x,h}(B \cap \mathrm{H}^c(v, c_{P,h}(v)))$. Note that P_1 is also a sub-measure of $P_{v_\infty,h}$. Set $P_2' = P_{v_\infty,h} - P_1$. Then, we have

$$\begin{aligned}
P_{x,h}\langle u, v\rangle &= P_1\langle u, v\rangle + P_2\langle u, v\rangle \\
&= P_{v_\infty,h}\langle u, v\rangle - P_2'\langle u, v\rangle + P_2\langle u, v\rangle \\
&= P_{v_\infty,h}\langle u, v\rangle - \langle m(P_2'), v\rangle P_2'(\mathbb{R}_d) + \langle m(P_2), v\rangle P_2(\mathbb{R}_d) \\
&= \bar{C}(P_{v_\infty,h}) - P_2'\langle u, v\rangle + P_2\langle u, v\rangle \\
&\leq \bar{C}(P_{v_\infty,h}) - c_{P,h}(v)P_2'(\mathbb{R}_d) + c_{P,h}(v)P_2(\mathbb{R}_d) \\
&= \bar{C}(P_{v_\infty,h}),
\end{aligned}$$

since P_2 is supported on $\mathrm{H}^c(v, c_{P,h}(v))$ and P_2' is supported on $\overline{\mathrm{H}}(v, c_{P,h}(v))$ and $P_2(\mathbb{R}_d) = P_2'(\mathbb{R}_d)$.

Thus, for all $x \in \mathbb{R}_d$, $\langle m(P_{x,h}), v\rangle \leq \bar{C}(P_{v_\infty,h}) = \langle m(P_{v_\infty,h}), v\rangle$. It means that $m(P_{v_\infty,h})$ is not included in a d-dimensional simplex within $\mathrm{Conv}(\mathcal{M}_h(P))$. It is thus a k-extreme point of $\mathrm{Conv}(\mathcal{M}_h(P))$ for some $k < d$. Moreover, the hyperplane $\partial H(v, c_{P,h}(v))$ separates $m(P_{v_\infty,h})$ from $\mathrm{Conv}(\mathcal{M}_h(P))$.

If x is extreme, then there is some vector v and some constant C_x such that $\langle m(P_{x,h}), v\rangle = C_x$ and such that for all y, $\langle m(P_{y,h}), v\rangle \leq C_x$. We aim at proving that x is in $\overline{\mathbb{R}}_d \backslash \mathbb{R}_d$. Similarly, we get

$$\begin{aligned}
C_x &= P_{x,h}\langle u, v\rangle \\
&= P_1\langle u, v\rangle + P_2\langle u, v\rangle \\
&= P_{v_\infty,h}\langle u, v\rangle - P_2'\langle u, v\rangle + P_2\langle u, v\rangle \\
&\leq C_x - P_2'\langle u, v\rangle + P_2\langle u, v\rangle \\
&\leq C_x - c_{P,h}(v)P_2'(\mathbb{R}_d) + c_{P,h}(v)P_2(\mathbb{R}_d) \\
&= C_x.
\end{aligned}$$

Thus the inequalities are equalities and we get that for P_2-almost all y, $\langle y, v\rangle = c_{P,h}(v)$ and for P_2'-almost all y, $\langle y, v\rangle = c_{P,h}(v)$. Thus, $P_{x,h}$ belongs to $\mathcal{P}_{v_\infty,h}(P)$. Note that, since there is equality, $C_x = \langle m(P_{v_\infty,h}), v\rangle = \langle m(P_{x,h}), v\rangle$.

Note that according to the Krein-Milman theorem, we get that $\mathrm{Conv}(\mathcal{M}_h(P)) = \mathrm{Conv}(\{m(P_{v_\infty,h}) \mid v \in \mathrm{S}(0,1)\})$.

We proved that for all $y \in \mathbb{R}_d$, for all $v \in \mathrm{S}(0,1)$,

$$\langle m(P_{-v_\infty,h}), v\rangle \leq \langle m(P_{y,h}), v\rangle \leq \langle m(P_{v_\infty,h}), v\rangle.$$

Therefore, the convex set $\mathrm{Conv}(\mathcal{M}_h(P))$ is included in $\bigcap_{v\in\mathrm{S}(0,1)} H^c(v, \langle m(P_{v_\infty,h}), v\rangle)$. With the Hahn-Banach separation theorem, we prove that for any $\theta \notin \mathrm{Conv}(\mathcal{M}_h(P))$, there is some vector v such that for all $\theta' \in \mathrm{Conv}(\mathcal{M}_h(P))$, $\langle \theta', v\rangle \leq C < \langle \theta, v\rangle$. In particular, we get that $\langle \theta, v\rangle > \langle m(P_{v_\infty,h}), v\rangle$, meaning that θ does not belong to $H(v, \langle m(P_{v_\infty,h}), v\rangle)$. ■

A.4 Proof of Lemma 5

Proof
Thanks to Lemma 4, we have:

$$\mathrm{Conv}(\mathcal{M}_h(P)) = \bigcap_{v\in\mathrm{S}(0,1)} \mathrm{H}^c(v, \langle m(P_{v_\infty,h}), v\rangle).$$

Let $0 < h' \leq h \leq 1$, in order to prove that the map $h \mapsto \mathrm{Conv}(\mathcal{M}_h(P))$ is non-increasing, it is sufficient to prove that

$$\mathrm{H}^c(v, \langle m(P_{v_\infty,h'}), v\rangle) \supset \mathrm{H}^c(v, \langle m(P_{v_\infty,h}), v\rangle).$$

Thus, it is sufficient to prove that

$$\langle m(P_{v_\infty,h'}), v\rangle \geq \langle m(P_{v_\infty,h}), v\rangle.$$

Set P_0 the sub-measure of P supported on $\overline{\mathrm{H}}^c(v, c_{P,h}(v))\backslash\mathrm{H}^c(v, c_{P,h'}(v))$ such that $hP_{v_\infty,h} = h'P_{v_\infty,h'} + (h-h')P_0$. Then, we have:

$$\langle m(P_{v_\infty,h}), v\rangle = \frac{h'}{h}\langle m(P_{v_\infty,h'}), v\rangle + \frac{h'-h}{h}\langle m(P_0), v\rangle.$$

The results comes from the fact that $\langle m(P_0), v\rangle \leq c_{P,h'}(v) \leq \langle m(P_{v_\infty,h'}), v\rangle$. ■

A.5 Proof of Lemma 7

The proof of Lemma 7 is based on the following concentration argument, that allows to connect empirical sub-measures with sub-measures for P_n. For sake of concision the statement also encompasses sub-Gaussian measures.

Lemma 25.
Suppose that $Q \in \mathcal{P}^{(V)}(\mathbb{R}_d)$. Then, for every $p > 0$, with probability larger than $1 - 8n^{-p}$,

we have,

$$\sup_{x,r} |(Q_n - Q)| \mathbb{1}_{\mathrm{B}(x,r)}(y)dy \leq C\sqrt{\frac{d+1}{n}} + \sqrt{\frac{2p\log(n)}{n}}$$

$$\sup_{v,t} |(Q_n - Q)| \mathbb{1}_{\langle y,v\rangle \leq t}(y)dy \leq C\sqrt{\frac{d+1}{n}} + \sqrt{\frac{2p\log(n)}{n}}$$

$$\sup_{x,r} \|(Q_n - Q) y \mathbb{1}_{\mathrm{B}(x,r)}(y)dy\| \leq CV\sqrt{d}\frac{(p+1)\log(n)}{\sqrt{n}}$$

$$\sup_{v,t} \|(Q_n - Q) y \mathbb{1}_{\langle v,y\rangle \leq t}dy\| \leq CV\sqrt{d}\frac{(p+1)\log(n)}{\sqrt{n}}$$

$$\sup_{x,r} \left|(Q_n - Q)\|y\|^2 \mathbb{1}_{\mathrm{B}(x,r)}(y)dy\right| \leq CV^2\sqrt{d}\frac{(p+1)\log(n)^{\frac{3}{2}}}{\sqrt{n}}$$

$$\sup_{v,t} \left|(Q_n - Q)\|y\|^2 \mathbb{1}_{\langle v,y\rangle \leq t}dy\right| \leq CV^2\sqrt{d}\frac{(p+1)\log(n)^{\frac{3}{2}}}{\sqrt{n}},$$

where $C > 0$ denotes a universal constant.

The proof of Lemma 25 is postponed to the following Section A.6. A significant part of the proof of Lemma 7 is based on the characterization of $\mathrm{Conv}(M_h(P))$ through $\omega^2_{P,h}$ stated by Lemma 16, where we recall that $\omega^2_{P,h}(\tau)$ is defined in Definition 11 by

$$\omega^2_{P,h}(\tau) = \sup_{x \in \mathbb{R}_d} \mathrm{d}^2_{P,h}(x) - \|x - \tau\|^2.$$

Lemma 26.
Let C denote a convex set, $\theta \in \mathbb{R}_d$, and $\Delta = \mathrm{d}(\theta, C)$. There exists $v \in \mathbb{R}_d$ with $\|v\| = 1$ such that, for all τ in C,

$$\langle v, \theta - \tau\rangle \geq \Delta.$$

Proof
[Proof of Lemma 26] Denote by π the projection onto C, and $t = \pi(\theta)$. Then, let $x = \frac{\theta - t}{\Delta}$. We may write

$$\begin{aligned}\langle x, \theta - \tau\rangle &= \langle x, \theta - t\rangle + \langle x, t - \tau\rangle \\ &= \Delta + \frac{1}{\Delta}\langle \theta - t, t - \tau\rangle.\end{aligned}$$

Since, for all τ in C, $\langle \theta - t, \tau - t\rangle \leq 0$, the result follows. ■

We are now in position to prove Lemma 7.

Proof
[Proof of Lemma 7] Let P in $\mathcal{P}^K(\mathbb{R}_d)$ which puts no mass on hyperplanes nor on spheres, and $\theta \in \mathrm{Conv}(\mathcal{M}_h(P))$. If we choose p large enough (for instance $p = 10$), a union bound ensures that the inequalities of Lemma 25 are satisfied for all $n \in \mathbb{N}$ with probability > 0. Since P puts no mass on hyperplanes, the probability that n points are not in general position is 0. Hence there exists an empirical distribution P_n, in general position, satisfying the inequalities of Lemma 25 for all n. In particular, for such a distribution P_n and (y, r) such that $P(\mathrm{B}(y, r)) = h$, we have

$$\left\| \frac{P_n u \mathbb{1}_{\mathrm{B}(y,r)}(u)}{P(\mathrm{B}(y,r))} - \frac{P_n u \mathbb{1}_{\mathrm{B}(y,r)}(u)}{P_n(\mathrm{B}(y,r))} \right\| \leq \frac{K\alpha_n}{h},$$

$$\left\| \frac{P u \mathbb{1}_{\mathrm{B}(y,r)}(u)}{P(\mathrm{B}(y,r))} - \frac{P_n u \mathbb{1}_{\mathrm{B}(y,r)}(u)}{P(\mathrm{B}(y,r))} \right\| \leq \frac{K\alpha_n}{h},$$

$$|(P_n - P)\mathrm{B}(y, r)| \leq \alpha_n,$$

for $\alpha_n \to 0$. Note that the same holds for means on half-spaces. Now let $x \in \mathbb{R}_d$,

$$\begin{aligned}
\mathrm{d}^2_{P,h}(x) - \|x-\theta\|^2 &= P_{x,h}\|x-u\|^2 - \|x-\theta\|^2 \\
&= \inf_{y\in\overline{\mathbb{R}}^d} P_{y,h}\|x-u\|^2 - \|x-\theta\|^2 \\
&\geq \inf_{y\in\overline{\mathbb{R}}^d} \|m(P_{y,h})\|^2 + v(P_{y,h}) - \|\theta\|^2 + \inf_{y\in\overline{\mathbb{R}}^d} 2\langle x, \theta - m(P_{y,h})\rangle \\
&\geq -\|\theta\|^2 + \inf_{y\in\overline{\mathbb{R}}^d} 2\langle x, \theta - m(P_{y,h})\rangle.
\end{aligned}$$

Thus, we may write

$$\begin{aligned}
\inf_{y\in\overline{\mathbb{R}}^d} 2\langle x, \theta - m(P_{y,h})\rangle &\geq \min\left[\inf_{y,r|P_n(\mathrm{B}(y,r))\in[h-\alpha_n,h+\alpha_n]} 2\left\langle x, \theta - \frac{P_n u \mathbb{1}_{\mathrm{B}_{y,r}(u)}}{P_n(\mathrm{B}(y,r))}\right\rangle,\right. \\
&\qquad \left.\inf_{v,t|P_n(H(v,t))\in[h-\alpha_n,h+\alpha_n]} 2\left\langle x, \theta - \frac{P_n u \mathbb{1}_{H(v,t)(u)}}{P_n(H(v,t))}\right\rangle\right] - \frac{4K\alpha_n\|x\|}{h}, \\
&= \inf_{\tau\in\bigcup_{s\in[h-\alpha_n,h+\alpha_n]}\mathcal{M}_s(P_n)} 2\langle x, \theta-\tau\rangle - \frac{4K\alpha_n\|x\|}{h}.
\end{aligned}$$

Now, if $\mathrm{d}\left(\theta, \mathrm{Conv}\left(\bigcup_{s\in[h-\alpha_n,h+\alpha_n]}\mathcal{M}_s(P_n)\right)\right) = \Delta > \frac{2K\alpha_n}{h}$, then according to Lemma 26, we can choose x in $\mathbb{R}_d$ such that, for all $\tau \in \mathrm{Conv}\left(\bigcup_{s\in[h-\alpha_n,h+\alpha_n]}\mathcal{M}_s(P_n)\right)$,

$$\left\langle \frac{x}{\|x\|}, \theta - \tau\right\rangle - \frac{2K\alpha_n}{h} > 0.$$

In this case, we immediately get $\omega^2_{P,h}(\theta) = \sup_{x\in\mathbb{R}_d} \mathrm{d}^2_{P,h}(x) - \|x-\theta\|^2 = +\infty$. According to Lemma 16, this contradicts $\omega \in \mathrm{Conv}(\mathcal{M}_h(P))$.

Set $h_n = \frac{q_n}{n}$ for $q_n \in [\![1, n]\!]$ such that $h - \alpha_n \geq h_n \geq h - \alpha_n - \frac{1}{n}$. Note that for n large enough, $h - \alpha_n - \frac{1}{n} > 0$, thus h_n is well defined. Then, according to Lemma 5 and 6,

$$\mathrm{Conv}\left(\bigcup_{s\in[h-\alpha_n,h+\alpha_n]}\mathcal{M}_s(P_n)\right) \subset \mathrm{Conv}\left(\bigcup_{s\in[h_n,1]}\mathcal{M}_s(P_n)\right) = \mathcal{M}_{\frac{q_n}{n}}(P_n).$$

Thus, we can build a sequence $(y_n)_{n\geq N}$ for some $N \in \mathbb{N}$ such that $y_n \in \mathcal{M}_{h_n}(P_n)$ and $\|\theta - y_n\| \leq 2\frac{K\alpha_n}{h}$. Hence the result of Lemma 7. ■

A.6 Proof of Lemma 25

Proof

[Proof of Lemma 25]

The first inequality is a direct application of Theorem 3.2 in [4], since the Vapnik dimension of balls in $\mathbb{R}^d$ is $d+1$. The same argument holds for the second inequality.

Now turn to the third one. Let $\lambda = p\log(n)$, $t = \sqrt{4V^2(\log(n)+\lambda)}$. Since $Q \in \mathcal{P}^{(V)}(\mathbb{R}_d)$, we have that

$$\mathbb{P}\left\{\max_i \|X_i\| \geq t\right\} \leq ne^{-\frac{t^2}{2V^2}} \leq n^{-2p+1}.$$

We may write

$$\begin{aligned}\sup_{x,r}\|(Q_n-Q)y\mathbb{1}_{\mathrm{B}(x,r)}(y)dy\| &= \sup_{x,r}\left\|\frac{1}{n}\sum_{i=1}^n X_i\mathbb{1}_{\mathrm{B}(x,r)}(X_i)-\mathbb{E}(X\mathbb{1}_{\mathrm{B}(x,r)}(X))\right\| \\ &\le \sup_{x,r}\left\|\frac{1}{n}\sum_{i=1}^n X_i\mathbb{1}_{\mathrm{B}(x,r)}(X_i)\mathbb{1}_{\|X_i\|\le t}-\mathbb{E}(X\mathbb{1}_{\mathrm{B}(x,r)}(X)\mathbb{1}_{\|X\|\le t})\right\| \\ &\quad + \mathbb{E}(\|X\|\mathbb{1}_{\|X\|>t})+\sup_{x,r}\frac{1}{n}\sum_{i=1}^n\|X_i\|\mathbb{1}_{\|X_i\|>t}.\end{aligned}$$

On one hand,

$$\begin{aligned}\mathbb{E}(\|X\|\mathbb{1}_{\|X\|>t}) &\le \sqrt{\mathbb{E}(\|X\|^2)}\sqrt{\mathbb{P}(\|X\|\ge t)} \\ &\le 2Ve^{-\frac{t^2}{4V^2}} \\ &\le 2Vn^{-(p+1)}.\end{aligned}$$

On the other hand, with probability larger than $1-n^{-2p+1}$, it holds

$$\sup_{x,r}\frac{1}{n}\sum_{i=1}^n\|X_i\|\mathbb{1}_{\|X_i\|>t}=0.$$

Now denote by $f_{x,r,v}$ the function $\left\langle y\mathbb{1}_{\mathrm{B}(x,r)}(y),v\right\rangle\mathbb{1}_{\|y\|\le t}$, for $v\in\mathrm{B}(0,1)$, so that

$$\sup_{x,r}\left\|\frac{1}{n}\sum_{i=1}^n X_i\mathbb{1}_{\mathrm{B}(x,r)}(X_i)\mathbb{1}_{\|X_i\|\le t}-\mathbb{E}(X\mathbb{1}_{\mathrm{B}(x,r)}(X)\mathbb{1}_{\|X\|\le t})\right\| = \sup_{x,r,v}|(Q_n-Q)f_{x,r,v}|.$$

A straightforward application of MacDiarmid's inequality (see, e.g., [5, Theorem 6.2]) entails

$$\mathbb{P}\left(\sup_{x,r,v}|(Q_n-Q)f_{x,r,v}|\ge\mathbb{E}\sup_{x,r,v}|(Q_n-Q)f_{x,r,v}|+t\sqrt{\frac{2\lambda}{n}}\right)\le e^{-\lambda}=n^{-p}.$$

It remains to bound $\mathbb{E}\sup_{x,r,v}|(Q_n-Q)f_{x,r,v}|$. A symmetrization inequality (see, e.g., [5, Lemma 11.4]) leads to

$$\mathbb{E}\sup_{x,r,v}|(Q_n-Q)f_{x,r,v}|\le\frac{2t}{n}\mathbb{E}_{X_{1:n}}\mathbb{E}_\varepsilon\sup_{x,r,v}\sum_{i=1}^n\varepsilon_i f_{x,r,v}(X_i)/t,$$

where the ε_i's are i.i.d. Rademacher random variable, and $\mathbb{E}_Y$ denotes expectation with respect to the random variable Y.

Now suppose that $X_1,\dots,X_n$ is fixed. In order to apply Dudley's entropy integral we have to provide an upper bound on the metric entropy of $\mathcal{F}=\{f_{x,r,v}/t\}_{x,r,v}$, for the $L_2(P_n)$ distance, that is $d^2(f,f')=\sum_{i=1}^n(f(X_i)-f'(X_i))^2/n$. Denote, for any subset of functions G, $\mathcal{N}(G,\varepsilon,L_2(P_n))$ the ε-covering number of G with respect to the metric $L_2(P_n)$. Then, if $G\subset G_1\times G_2$ (for the multiplication), with $\|G_2\|\le 1$ and $\|G_1\|\le 1$, we may write

$$\mathcal{N}(G,\varepsilon,L_2(P_n))\le\mathcal{N}(G_1,\varepsilon/(2\sqrt{2}),L_2(P_n))\times\mathcal{N}(G_2,\varepsilon/(2\sqrt{2}),L_2(P_n)).$$

Define $G_1=\{(\langle v,.\rangle)/t\}_{\|v\|\le 1}$, and $G_2=\left\{\mathbb{1}_{\mathrm{B}(x,r)\cap B(0,t)}\right\}_{x,r}$. It is obvious that $\mathcal{F}\subset G_1\times G_2$. Using Theorem 1 in [11], if $\|G\|\le 1$, we have

$$\mathcal{N}(G,\varepsilon,L_2(P_n))\le\left(\frac{2}{\varepsilon}\right)^{Cd_p(G)},$$

where C is an absolute constant and d_p denotes the pseudo-dimension. Hence we have

$$\mathcal{N}(G_1, \varepsilon, L_2(P_n)) \leq \left(\frac{2}{\varepsilon}\right)^{Cd}$$
$$\mathcal{N}(G_2, \varepsilon, L_2(P_n)) \leq \left(\frac{2}{\varepsilon}\right)^{C(2(d+1))}.$$

We may then deduce

$$\mathcal{N}(\mathcal{F}, \varepsilon, L_2(P_n)) \leq \left(\frac{4\sqrt{2}}{\varepsilon}\right)^{C(3d+2)}.$$

Now, using Dudley's entropy integral (see, e.g., [5, Corollary 13.2]) yields

$$\begin{aligned}\mathbb{E}_\varepsilon \frac{1}{\sqrt{n}} \sup_{x,r,v} \sum_{i=1}^n \varepsilon_i f_{x,r,v}(X_i)/t &\leq 12 \int_0^1 \sqrt{\log(\mathcal{N}(\mathcal{F}, \varepsilon, L_2(P_n)} d\varepsilon \\ &\leq 12\sqrt{C(3d+2)} \int_0^1 \sqrt{\log(\frac{4\sqrt{2}}{\varepsilon})} d\varepsilon \\ &\leq C\sqrt{d}.\end{aligned}$$

Hence we deduce that

$$\mathbb{E} \sup_{x,r,v} |(Q_n - Q) f_{x,r,v}| \leq \frac{Ct\sqrt{d}}{\sqrt{n}} \leq \frac{CV\sqrt{(p+1)\log(n)}}{\sqrt{n}}.$$

Combining the different terms gives, with probability larger than $1 - 2n^{-p}$,

$$\sup_{x,r} \|(Q_n - Q) y \mathbb{1}_{\mathrm{B}(x,r)}(y) dy\| \leq CV \frac{(p+1)\log(n)}{\sqrt{n}}.$$

the third deviation bound follows. The fourth deviation bound may be proved the same way.

For the 5-th inequality, as before let $\lambda = p\log(n)$ and $t = \sqrt{4V^2(\log(n) + \lambda)}$. Similarly, we may write

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

with $\mathbb{E}(\|X\|^2 \mathbb{1}_{\|X\|>t}) \leq 2V^2 n^{-(p+1)}$ and, with probability larger than $1 - n^{-\frac{1}{p}}$,

$$\sup_{x,r} \frac{1}{n} \sum_{i=1}^n \|X_i\|^2 \mathbb{1}_{\|X_i\| > t} = 0.$$

Using [5, Theorem 6.2] again leads to

$$\begin{aligned}&\mathbb{P}\left(\sup_{x,r} \left| \frac{1}{n} \sum_{i=1}^n \|X_i\|^2 \mathbb{1}_{\mathrm{B}(x,r)}(X_i) \mathbb{1}_{\|X_i\| \leq t} - \mathbb{E}(\|X\|^2 \mathbb{1}_{\mathrm{B}(x,r)}(X) \mathbb{1}_{\|X\| \leq t}) \right| \right. \\ &\geq \left. \mathbb{E} \sup_{x,r} \left| \frac{1}{n} \sum_{i=1}^n \|X_i\|^2 \mathbb{1}_{\mathrm{B}(x,r)}(X_i) \mathbb{1}_{\|X_i\| \leq t} - \mathbb{E}(\|X\|^2 \mathbb{1}_{\mathrm{B}(x,r)}(X) \mathbb{1}_{\|X\| \leq t}) \right| + t^2 \sqrt{\frac{2\lambda}{n}} \right) \leq e^{-\lambda} = n^{-\frac{1}{p}}.\end{aligned}$$

At last, combining a symmetrization inequality with a contraction principle ([5, Theorem 11.5]) gives

$$\begin{aligned}
&\mathbb{E}\sup_{x,r}\left|\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2\mathbb{1}_{\mathrm{B}(x,r)}(X_i)\mathbb{1}_{\|X_i\|\leq t}-\mathbb{E}(\|X\|^2\mathbb{1}_{\mathrm{B}(x,r)}(X)\mathbb{1}_{\|X\|\leq t})\right| \\
&\leq \frac{2t^2}{n}\mathbb{E}_{X_{1:n}}\mathbb{E}_{\varepsilon}\sup_{x,r}\sum_{i=1}^{n}\varepsilon_i\frac{\|X_i\|^2}{t^2}\mathbb{1}_{\mathrm{B}(x,r)\cap\mathrm{B}(0,t)}(X_i) \\
&\leq \frac{2t^2}{n}\mathbb{E}_{X_{1:n}}\mathbb{E}_{\varepsilon}\sup_{x,r}\sum_{i=1}^{n}\varepsilon_i\mathbb{1}_{\mathrm{B}(x,r)\cap\mathrm{B}(0,t)}(X_i) \\
&\leq \frac{C\sqrt{d}t^2}{\sqrt{n}},
\end{aligned}$$

where the last line may be derived the same way as for the third inequality, combining [11, Theorem 1] and [5, Corollary 13.2]. Gluing all pieces yields, with probability larger than $1-2n^{-\frac{1}{p}}$,

$$\sup_{x,r}\left|(Q_n-Q)\|y\|^2\mathbb{1}_{\mathrm{B}(x,r)}(y)dy\right|\leq CV^2\sqrt{d}\frac{(p+1)\log(n)^{\frac{3}{2}}}{\sqrt{n}}.$$

The last inequality follows from the same argument. ■

A.7 Proof of Lemma 10

Proof
In Proposition 3.6 from [7], we get:

$$\begin{aligned}
\mathrm{d}^2_{P,h}(y) &= P_{y,h}\|y-u\|^2 \\
&\leq P_{x,h}\|y-u\|^2 \\
&= \|y-x\|^2+P_{x,h}\|x-u\|^2+2\langle y-x,x-P_{x,h}u\rangle.
\end{aligned}$$

In particular,

$$\mathrm{d}^2_{P,h}(y)-\|y\|^2\leq \mathrm{d}^2_{P,h}(x)-\|x\|^2-2\langle y-x,P_{x,h}u\rangle,$$

with equality if and only if $P_{y,h}\|y-u\|^2=P_{x,h}\|y-u\|^2$, that is if and only if like $hP_{y,h}$, $hP_{x,h}$ is also a sub-measure of P with total mass h, whose support is contained in the closed ball $\overline{\mathrm{B}}(y,\delta_{P,h}(y))$ and whose restriction to the open ball $\mathrm{B}(y,\delta_{P,h}(y))$ coincides with P; see [7], Proposition 3.3. ■

B Proofs for Section 3

B.1 Proof of Theorem 13

Proof
First note that for all $t,s\in\overline{\mathbb{R}}_d^{(k)}$, denoting by

$$f_s:x\mapsto\min_{i\in[\![1,k]\!]}\left(\|x-m(P_{s_i,h})\|^2+v(P_{s_i,h})\right),$$

we have:

$$Pf_s(u)-f_t(u)\leq\sum_{i=1}^{k}\tilde{P}_{t_i,h}(\mathbb{R}_d)\left(2\langle\tilde{m}(\tilde{P}_{t_i,h}),m(P_{t_i,h})-m(P_{s_i,h})\rangle+M(P_{s_i,h})-M(P_{t_i,h})\right).$$

Then, according to Lemma 3 and the dominated convergence Theorem, for any $v \in \mathrm{S}(0, 1)$, there is a sequence $(x_n)_{n\in\mathbb{N}}$ in $\mathbb{R}_d$ such that $m(P_{x_n,h}) \to m(P_{v_\infty,h})$ and $M(P_{x_n,h}) \to M(P_{v_\infty,h})$. Then, $\limsup_{n\to+\infty} Pf_{x_n}(u) - f_{v_\infty}(u) \leq 0$. Thus, $\inf_{t\in\mathbb{R}_d} Pf_t(u) = \inf_{t\in\overline{\mathbb{R}}_d} Pf_t(u)$.

Let $(t_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}_d^{(k)}$ such that $Pf_{t_n}(u) \leq \inf_{t\in\overline{\mathbb{R}}_d} Pf_t(u) + \frac{1}{n}$, and denote by m^* the limit of a converging subsequence of $\big(\tilde{m}(\tilde{P}_{t_{n,1},h}), \tilde{m}(\tilde{P}_{t_{n,2},h}), \ldots, \tilde{m}(\tilde{P}_{t_{n,k},h})\big)_{n\in\mathbb{N}}$ in the compact space $\overline{\mathrm{B}}(0, K)^{(k)}$. Then, thanks to Lemma 10, and recalling that $\forall y \in \mathbb{R}_d$, $\mathrm{d}^2_{P,h}(y) = \|y - m(P_{y,h})\|^2 + v(P_{y,h})$,

$$Pf_{m^*}(u) - f_{t_n}(u) \leq 2\sum_{i=1}^{k} \tilde{P}_{t_{n,i},h}(\mathbb{R}_d)\langle \tilde{m}(\tilde{P}_{t_{n,i},h}) - m_i^*, m(P_{t_{n,i},h}) - m(P_{m_i^*,h})\rangle,$$

which goes to zero when $n \to +\infty$ since $\|m(P_{y,h})\| \leq K$ whenever $y \in \mathbb{R}_d$. Thus, $Pf_{m^*}(u) = \inf_{t\in\overline{\mathbb{R}}_d} Pf_t(u)$. In particular, there is some $s \in \mathcal{O}pt(P, h, k) \cap \overline{\mathrm{B}}(0, K)^{(k)}$. Take $P_{s_i,h} \in \mathcal{P}_{s_i,h}(P)$ for all $i \in [\![1, k]\!]$, then set $s^* \in \overline{\mathrm{B}}(0, K)^{(k)}$ such that $s_i^* = \tilde{m}(\tilde{P}_{s_i,h})$ for all $i \in [\![1, k]\!]$. Then for any choice of $P_{s_i^*,h} \in \mathcal{P}_{s_i^*,h}(P)$,

$$\begin{aligned}
0 &\leq Pf_{s^*}(u) - f_s(u) \\
&\leq \sum_{i=1}^{k} \tilde{P}_{s_i,h}\|u - m(P_{s_i^*,h})\|^2 + v(P_{s_i^*,h}) - \|u - m(P_{s_i,h})\|^2 - v(P_{s_i,h}) \\
&= \sum_{i=1}^{k} \tilde{P}_{s_i,h}\left(\mathrm{d}^2_{P,h}(s_i^*) - \|s_i^*\|^2\right) - \left(\mathrm{d}^2_{P,h}(s_i) - \|s_i\|^2\right) - \langle u - s_i^*, 2m(P_{s_i^*,h})\rangle + \langle u - s_i, 2m(P_{s_i,h})\rangle \\
&\leq 2\sum_{i=1}^{k} \tilde{P}_{s_i,h}(\mathbb{R}_d)\langle \tilde{m}(\tilde{P}_{s_i,h}) - s_i^*, m(P_{s_i,h}) - m(P_{s_i^*,h})\rangle = 0.
\end{aligned}$$

Thus, inequalities are all equalities. In particular, equality in Lemma 10 leads to $P_{s_i,h} \in \mathcal{P}_{s_i^*,h}(P)$, and by choosing $P_{s_i^*,h} = P_{s_i,h}$, the Laguerre measures $(\tilde{P}_{s_i,h})_{i\in[\![1,k]\!]}$ are also appropriate for s^*. Then, $\tilde{m}(\tilde{P}_{s_i^*,h}) = \tilde{m}(\tilde{P}_{s_i,h}) = s_i^*$. Thus, $s^* \in \mathcal{O}pt(P, h, k) \cap \overline{\mathrm{B}}(0, K)^{(k)}$ and satisfies for some $(P_{s_i^*,h})_{i\in[\![1,k]\!]}$, $\tilde{m}(\tilde{P}_{s_i^*}) = s_i^*$, for all $i \in [\![1, k]\!]$. ■

B.2 Proof of Corollary 19

Proof

[Proof of Corollary 19] The proof of Corollary 19 is based on the following bounds, in the case where P is absolutely continuous with respect to the Lebesgue measure, with density f satisfying $0 < f_{min} \leq f \leq f_{max}$.

$$f_M^{-1}(k) \leq 2K\sqrt{d}k^{-1/d} \tag{6}$$

$$\zeta_{P,h}(f_M^{-1}(k)) \leq KC_{f_{max},K,d,h}k^{-1/d}. \tag{7}$$

The first equation proceeds from the following. Since $M \subset \mathrm{B}(0, K)$, for any $\varepsilon > 0$ we have

$$f_M(\varepsilon) \leq f_{\mathrm{B}(0,K)}(\varepsilon) \leq \left(\frac{2K\sqrt{d}}{\varepsilon}\right)^d.$$

Hence (6). To prove the second inequality, we will use the following Lemma.

Lemma 27.
Suppose that P has a density f satisfying $0 < f_{min} \leq f \leq f_{max}$. Let x, y be in M, and denote by $\delta = \|x - y\|$. Then

$$\|m(P_{x,h}) - m(P_{y,h})\| \leq (2K)^{d+1}\omega_d \left(1 + \delta \left(\frac{f_{max}\omega_d}{h}\right)^{1/d}\right)^{d-1} \left(\frac{f_{max}\omega_d}{h}\right)^{1/d} \delta.$$

Proof
[Proof of Lemma 27] Since P has a density, $P\partial \mathrm{B}(x, \delta_{x,h}) = P\partial \mathrm{B}(y, \delta_{y,h}) = 0$. We deduce that $P_{x,h} = \frac{1}{h} P_{|\mathrm{B}(x,h)}$ and $P_{y,h} = \frac{1}{h} P_{|\mathrm{B}(y,h)}$. Without loss of generality, assume that $\delta_{x,h} \geq \delta_{y,h}$. Then $\mathrm{B}(y, \delta_{y,h}) \subset \mathrm{B}(x, \delta_{x,h} + \delta)$. We may write

$$\begin{aligned}
\|m(P_{x,h}) - m(P_{y,h})\| &= \frac{1}{m_0} \left\| P\left(u(\mathbb{1}_{\mathrm{B}(x,\delta_{x,h})}(u) - \mathbb{1}_{\mathrm{B}(y,\delta_{y,h})}(u)\right)\right\| \\
&\leq \frac{2K}{h} P\left|(\mathbb{1}_{\mathrm{B}(x,\delta_{x,h})}(u) - \mathbb{1}_{\mathrm{B}(y,\delta_{y,h})}(u)\right| \\
&\leq \frac{2K}{h} P\left(\mathrm{B}(x, \delta_{x,h} + \delta) \cap \mathrm{B}(x, \delta_{x,h} + \delta)^c\right) \\
&\leq \frac{2K}{h} \omega_d \left[(\delta_{x,h} + \delta)^d - \delta_{x,h}^d\right] \\
&\leq \frac{(2K)^{d+1}\omega_d}{h} \left[(1 + \frac{\delta}{\delta_{x,h}})^d - 1\right].
\end{aligned}$$

Since $(1+v)^d \leq 1 + d(1+v)^{d-1}v$, for $v \geq 0$, and $\delta_{x,h} \geq \left(\frac{h}{f_{max}\omega_d}\right)^{1/d}$, the result follows. ■

Hence (7). The result of Corollary 19 follows. ■

B.3 Proof of Corollary 20

Proof
[Proof of Corollary 20] 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 $P(N^o) = 1$. Thus we take $M = N^o$, that is the set of interior points. Since P satisfies a (cf_{min}, d')-standard assumption, we have

$$f_M(\varepsilon) \leq \frac{2^{d'}}{cf_{min}} r^{-d'},$$

according to [8, 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 $P_{x,h} = P_{|\mathrm{B}(x,h)}$. Besides, since for all $r > 0$ $P(\mathrm{B}(x,r)) \geq cf_{min} r^{d'}$, we may write $\delta_{x,h} \leq c_{N,f_{min}} h^{1/d'} \leq \rho/12$, 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 $\delta_{x,h} \geq \delta_{y,h}$. Then, proceeding as in the proof of Lemma 27, it comes

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

Since $\delta_{x,h} + \delta \leq \rho/6$, for any u in $\mathrm{B}(x, \delta_{x,h} + \delta) \cap M$ 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., [1, Proposition 25]). Note that, according to [1, Proposition 26], for any u_1 and u_2 such that $\|u_1 - u_2\| \leq \rho/4$,

$$\|u_1 - u_2\| \leq d_N(u_1, u_2) \leq 2\|u_1 - u_2\|. \tag{8}$$

Now let $p_1, \dots, p_m$ be a δ-covering set of the sphere $\mathcal{S}_{x,\delta_{x,h}} = \{u \in M | \|x-u\| = \delta_{x,u}\}$. According to (8), we may choose $m \leq c_{d'} \delta_{x,h}^{d'-1} \delta^{-(d'-1)}$.

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

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

Now, for any j, since $2\delta \leq \rho/6$, in local polar coordinates around p_j we may write, using (8) again,

$$\begin{aligned} P\left(\mathrm{B}(p_j, 2\delta)\right) &\leq \int_{r,v|\exp_{p_j}(rv)\in M, r\leq 4\delta} f(r,v)J(r,v)drdv \\ &\leq f_{max} \int_{r,v|r\leq 4\delta} J(r,v)drdv \end{aligned}$$

where $J(r,v)$ denotes the Jacobian of the volume form. According to [1, Proposition 27], we have $J(r,v) \leq C_{d'} r^{d'}$. Hence $P\left(\mathrm{B}(p_j, 2\delta)\right) \leq C_{d'} f_{max} \delta^{d'}$. We may conclude

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

Choosing k large enough so that $f_M^{-1}(k) \leq C_{f_{min},N} k^{-1/d'} \leq \rho/12$ gives the result of Corollary 20. ■

B.4 Proof of Proposition 21

Proof

To lighten the notation we omit the ε in $\mathrm{d}_{Q,h,k,\varepsilon}$. For all $x \in \mathrm{Supp}(P)$,

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

Thus, $\left(\mathrm{d}^2_{Q,h,k} - \mathrm{d}^2_{P,h}\right)_- \leq \|\mathrm{d}^2_{P,h} - \mathrm{d}^2_{Q,h}\|_{\infty,\mathrm{Supp}(P)}$ on $\mathrm{Supp}(P)$, 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,

$$\begin{aligned} P\left|\mathrm{d}^2_{Q,h,k} - \mathrm{d}^2_{P,h}\right|(u) &= P\mathrm{d}^2_{Q,h,k}(u) - \mathrm{d}^2_{P,h}(u) + 2\left(\mathrm{d}^2_{Q,h,k}(u) - \mathrm{d}^2_{P,h}(u)\right)_- \\ &\leq P\Delta(u) + P\mathrm{d}^2_{P,h,k}(u) - \mathrm{d}^2_{P,h}(u) + 2\|\mathrm{d}^2_{P,h} - \mathrm{d}^2_{Q,h}\|_{\infty,\mathrm{Supp}(P)}. \end{aligned}$$

with $\Delta = \mathrm{d}^2_{Q,h,k} - \mathrm{d}^2_{P,h,k}$. We can bound $P\Delta(u)$ from above. Let $s \in \mathcal{O}pt(P,h,k) \cap \overline{\mathrm{B}}(0,K)^{(k)}$ such that $s_i = \tilde{m}(\tilde{P}_{s_i,h})$ for all $i \in [\![1,k]\!]$. Such an s exists according to Theorem 13. Set $f_{Q,t}(x) = 2\langle x, m(Q_{t,h})\rangle + v(Q_{t,h})$ for $t \in \mathbb{R}_d$, and let $t \in \mathcal{O}pt(Q,h,k)$.

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

For any transport plan π between P and Q,

$$\begin{aligned} P - Q \min_{i\in[\![1,k]\!]} f_{Q,t_i}(u) &= \mathbb{E}_{(X,Y)\sim\pi}\left[\min_{i\in[\![1,k]\!]} 2\langle X, m(Q_{t,h})\rangle + v(Q_{t,h}) - \min_{i\in[\![1,k]\!]} 2\langle Y, m(Q_{t,h})\rangle + v(Q_{t,h})\right] \\ &\leq 2\mathbb{E}_{(X,Y)\sim\pi}\left[\sup_{t\in\overline{\mathbb{R}}_d} \langle X-Y, m(Q_{t,h})\rangle\right]. \end{aligned}$$

Thus, $P - Q \min_{i\in[\![1,k]\!]} f_{Q,t_i}(u) \leq 2W_1(P,Q)\sup_{t\in\overline{\mathbb{R}}_d} m(Q_{t,h})$, after taking for π the optimal transport plan for the L_1-Wasserstein distance (noted W_1) between P and Q.

Also note that $P\min_{i\in[\![1,k]\!]} f_{Q,t_i}(u) - \min_{i\in[\![1,k]\!]} f_{P,s_i}(u)$ is bounded from above by

$$\begin{aligned}
&\leq \sum_{i=1}^{k} \tilde{P}_{s_i}\left(2\langle u, m(Q_{s_i,h})\rangle + v(Q_{s_i,h})\right) - \left(2\langle u, m(P_{s_i,h})\rangle + v(P_{s_i,h})\right)\\
&= \sum_{i=1}^{k} \tilde{P}_{s_i} \min_{j\in[\![1,k]\!]} \left(2\langle u - s_i, m(P_{s_i,h}) - m(Q_{s_i,h})\rangle + \mathrm{d}^2_{Q,h}(s_i) - \mathrm{d}^2_{P,h}(s_i)\right)\\
&\leq \|\mathrm{d}^2_{P,h} - \mathrm{d}^2_{Q,h}\|_{\infty,\mathrm{Supp}(P)} + 2\sum_{i=1}^{k} \tilde{P}_{s_i,h}(\mathbb{R}_d)\langle \tilde{m}(\tilde{P}_{s_i,h}) - s_i, m(P_{s_i,h}) - m(Q_{s_i,h})\rangle.
\end{aligned}$$

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

■

B.5 Proof of Proposition 22

Proof

The proof of Proposition 22 relies on [7, Corollary 4.8]. Namely, if P satisfies (5), then

$$\|d_{P,h} - d_M\|_\infty \leq C(P)h^{-\frac{1}{d'}}.$$

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

$$\Delta_P^2 \geq C(P)\Delta_\infty^{d'+2}.$$

Thus we have $\sup_{x\in M} |d_{Q,h,k} - d_M|(x) \leq C(P)^{-\frac{1}{d'+2}}\Delta_P^{\frac{2}{d'+2}}$. 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 $d_{Q,h,k,\varepsilon}(p) = \sqrt{\|p - m(Q_{t_j,h})\|^2 + v(Q_{t_j,h})}$. Then

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

On the other hand, we have $d_{Q,h,k,\varepsilon} \geq d_{Q,h}$, along with $\|d_{Q,h} - d_{P,h}\|_\infty \leq h^{-\frac{1}{2}}W_2(P,Q)$ (see, e.g., [7, Theorem 3.5]) as well as $d_{P,h} \geq d_M$. Hence

$$d_{Q,h,k,\varepsilon} \geq d_M - h^{-\frac{1}{2}}W_2(P,Q).$$

■

C Proofs for Section 4

C.1 Proof of Proposition 23

Proof

For any $t = (t_1, t_2, \dots t_k) \in \mathbb{R}_d^{(k)}$, we note $c_i = \frac{\sum_{X \in \mathcal{C}(t_i)} X}{|\mathcal{C}(t_i)|}$. Then,

$$
\begin{aligned}
& P_n \min_{i \in [\![1,k]\!]} \|u - m(P_{n\,t_i,h})\|^2 + v(P_{n\,t_i,h}) \\
&= \sum_{i=1}^{k} \frac{1}{n} \sum_{X \in \mathcal{C}(t_i)} \|X - m(P_{n\,t_i,h})\|^2 + v(P_{n\,t_i,h}) \\
&= \sum_{i=1}^{k} \frac{1}{n} \sum_{X \in \mathcal{C}(t_i)} \|X\|^2 - 2\langle X - t_i, m(P_{n\,t_i,h})\rangle + \left(\mathrm{d}^2_{P_n,h}(t_i) - \|t_i\|^2\right) \\
&= \frac{1}{n} \sum_{j=1}^{n} \|X_j\|^2 + \sum_{i=1}^{k} \frac{|\mathcal{C}(t_i)|}{n} \left(-2\langle c_i - t_i, m(P_{n\,t_i,h})\rangle + \left(\mathrm{d}^2_{P_n,h}(t_i) - \|t_i\|^2\right)\right) \\
&\geq \frac{1}{n} \sum_{j=1}^{n} \|X_j\|^2 + \sum_{i=1}^{k} \mathrm{d}^2_{P_n,h}(c_i) - \|c_i\|^2 \\
&= \sum_{i=1}^{k} \frac{1}{n} \sum_{X \in \mathcal{C}(t_i)} \|X - m(P_{n\,c_i,h})\|^2 + v(P_{n\,c_i,h}) \\
&\geq \sum_{i=1}^{k} \frac{1}{n} \sum_{X \in \mathcal{C}(c_i)} \|X - m(P_{n\,c_i,h})\|^2 + v(P_{n\,c_i,h}) \\
&= P_n \min_{i \in [\![1,k]\!]} \|u - m(P_{n\,c_i,h})\|^2 + v(P_{n\,c_i,h}).
\end{aligned}
$$

We used Lemma 10. ■

C.2 Proof of Theorem 24

Let γ and $\hat{\gamma}$ the functions defined for $(t, x) \in \mathbb{R}_d^{(k)} \times \mathbb{R}_d$ with $t = (t_1, t_2, \dots, t_k)$, by:

$$\gamma(t, x) = \min_{i \in [\![1,k]\!]} -2\langle x, m(Q_{t_i,h})\rangle + \|m(Q_{t_i,h})\|^2 + v(Q_{t_i,h}),$$

and

$$\hat{\gamma}(t, x) = \min_{i \in [\![1,k]\!]} -2\langle x, m(Q_{n\,t_i,h})\rangle + \|m(Q_{n\,t_i,h})\|^2 + v(Q_{n\,t_i,h}).$$

The proof of Theorem 24 is based on the two following deviation Lemmas.

Lemma 28.
If Q is sub-Gaussian with variance V^2, then, for every $p > 0$, with probability larger than $1 - 2n^{-\frac{1}{p}}$, we have

$$\sup_{t \in \mathbb{R}_d^{(k)}} |(Q - Q_n)\gamma(t, u)| \leq C \frac{\sqrt{kd} V^2 \log(n)}{h\sqrt{n}}.$$

The proof of Lemma 28 is deferred to Section C.3.

Lemma 29.
Assume that Q is sub-Gaussian with variance V^2, then, for every $p > 0$, with probability larger than $1 - 7n^{-p}$, we have

$$\sup_{t \in \mathbb{R}_d^{(k)}} |Q_n(\gamma - \hat{\gamma})(t, u)| \leq CV^2 \frac{(p+1)^{\frac{3}{2}} \log(n)^{\frac{3}{2}}}{h\sqrt{n}}.$$

As well, the proof of Lemma 29 is deferred to Section C.4. We are now in position to prove Theorem 24.

Proof
[Proof of Theorem 24] Let

$$s = \arg\min \left\{ Q\gamma(t, u) \mid t = (t_1, t_2, \dots t_k) \in \mathbb{R}_d^{(k)} \right\},$$

$$\hat{s} = \arg\min \left\{ Q_n\hat{\gamma}(t, u) \mid t = (t_1, t_2, \dots t_k) \in \mathbb{R}_d^{(k)} \right\}$$

and

$$\tilde{s} = \arg\min \left\{ Q_n\gamma(t, u) \mid t = (t_1, t_2, \dots t_k) \in \mathbb{R}_d^{(k)} \right\}.$$

With these notations, for all $x \in \mathbb{R}_d$, $\mathrm{d}^2_{Q,h,k}(x) = \|x\|^2 + \gamma(s, x)$ and $\mathrm{d}^2_{Q_n,h,k}(x) = \|x\|^2 + \hat{\gamma}(\hat{s}, x)$. We intend to bound $l(s, \hat{s}) = Q(\mathrm{d}^2_{Q_n,h,k}(u) - \mathrm{d}^2_{Q,h,k}(u))$, which is also equal to $l(s, \hat{s}) = Q(\gamma(\hat{s}, u) - Q\gamma(s, u))$.

We have that:

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

where we used the fact that $Q_n\gamma(\tilde{s}, u) \leq Q_n\gamma(s, u)$. Now, since $Q_n(\hat{\gamma}(\hat{s}, u) - \hat{\gamma}(\tilde{s}, u)) \leq 0$, we get:

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

Combining Lemma 28 and Lemma 29 entails, with probability larger than $1 - 8n^{-p}$,

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

It remains to bound $|P\mathrm{d}^2_{Q_n,h,k} - Q\mathrm{d}^2_{Q_n,h,k}|$ as well as $|P\mathrm{d}^2_{Q,h,k} - Q\mathrm{d}^2_{Q,h,k}|$. To this aim we recall that $X = Y + Z$, Z being sub-Gaussian with variance σ^2. Thus, denoting by $t_j(x) = \arg\min_j \|x - m(Q_{t_j,h})\|^2 + v(Q_{t_j,h})$,

$$\begin{aligned}
P\mathrm{d}^2_{Q,h,k} - Q\mathrm{d}^2_{Q,h,k} &\leq \mathbb{E}\left[\|Y - m(Q_{t_j(Y),h})\|^2 + v(Q_{t_j(Y),h}) - \left(\|Y + Z - m(Q_{t_j(Y),h})\|^2 + v(Q_{t_j(Y),h}) \right) \right] \\
&\leq \mathbb{E}\|Z\|^2 + 2\mathbb{E} \max_{j \in [\![1,k]\!]} \left| \left\langle Z, m(Q_{t_j,h}) - (Y + Z) \right\rangle \right| \\
&\leq \sigma^2 + 2\sigma(\max_{j \in [\![1,k]\!]} \|m(Q_{t_j,h})\| + \sqrt{2}(K + \sigma)) \\
&\leq \frac{C\sigma K}{\sqrt{h}},
\end{aligned}$$

using (9) and $\sigma \leq K$. The converse bound on $P\mathrm{d}^2_{Q,h,k} - Q\mathrm{d}^2_{Q,h,k}$ may be proved the same way. Similarly, we may write

$$\begin{aligned} P\mathrm{d}^2_{Q_n,h,k} - Q\mathrm{d}^2_{Q_n,h,k} &\leq \sigma^2 + 2\sigma(\max_{j\in[\![1,k]\!]} \|m(Q_{n,t_j,h})\| + \sqrt{2}(K+\sigma) \\ &\leq \sigma^2 + 2\sigma(\max_{j\in[\![1,k]\!]} \|m(Q_{t_j,h})\| + \frac{C(K+\sigma)(p+1)\log(n)}{h\sqrt{n}}) + \sqrt{2}(K+\sigma)) \\ &\leq \frac{C\sigma K(p+1)\log(n)}{h\sqrt{n}}, \end{aligned}$$

according to Lemma 2. The bound on $Q\mathrm{d}^2_{Q_n,h,k} - P\mathrm{d}^2_{Q_n,h,k}$ derives from the same argument. Collecting all pieces, we have

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

where we used $\sigma \leq K$.

■

C.3 Proof of Lemma 28

Proof

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:

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

First we note that since Q is sub-Gaussian with variance V^2, we have, for every $c \in \overline{\mathbb{R}}_d$,

$$\|m(Q_{c,h})\|^2 + v(Q_{c,h}) = Q_{t,h}\|u\|^2 \leq \frac{2V^2}{h}. \tag{9}$$

Set $z = 2V\sqrt{\log(n)+\lambda}$ and $\lambda = p\log(n)$. Then, with probability larger than $1 - n^{-\frac{1}{p}}$,

$$\max_{i=1,\ldots,n} \|X_i\| \leq z. \tag{10}$$

We may then write

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

According to (10), the last part is 0 with probability larger than $1 - n^{-\frac{1}{p}}$. Moreover

$$\begin{aligned}
\mathbb{E}(|\gamma(t,X)|\mathbb{1}_{\|X\|>z}) &\leq \mathbb{E}(\mathbb{1}_{\|X\|>z} \sup_{j=1,\dots,k} 2|\langle x, m(Q_{t_j,h})|\rangle| + \|m(Q_{t_j,h})\|^2 + v(Q_{t_j,h}) \\
&\leq \frac{2V^2}{h}\mathbb{P}(\|X\| > z) + 2\sqrt{2}\frac{V}{\sqrt{h}}\mathbb{E}(\|X\|\mathbb{1}_{\|X\|>z}) \\
&\leq 10\frac{V^2}{h}e^{-\frac{z^2}{2V^2}} \\
&\leq 10\frac{V^2}{h}n^{-(p+1)}.
\end{aligned}$$

It remains to bound

$$\sup_{t\in\mathbb{R}_d^{(k)}} \left|(Q-Q_n)\gamma(t,u)\mathbb{1}_{\|u\|\leq z}\right|.$$

Since for every t and u, $|\gamma(t,u)\mathbb{1}_{\|u\|\leq z}| \leq (z + \frac{V\sqrt{2}}{\sqrt{h}})^2 := R$, [5, Theorem 6.2] entails

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}_d^{(k)}} \left|(Q-Q_n)\gamma(t,u)\mathbb{1}_{\|u\|\leq z}\right| \geq \mathbb{E}\sup_{t\in\mathbb{R}_d^{(k)}} \left|(Q-Q_n)\gamma(t,u)\mathbb{1}_{\|u\|\leq z}\right| + R\sqrt{\frac{2\lambda}{n}}\right) \leq e^{-\lambda} = n^{-p}.$$

To bound $\mathbb{E}\sup_{t\in\mathbb{R}_d^{(k)}} \left|(Q-Q_n)\gamma(t,u)\mathbb{1}_{\|u\|\leq z}\right|$, we follow the same line as for Lemma 25. A symmetrization argument yields

$$\begin{aligned}
\mathbb{E}\sup_{t\in\mathbb{R}_d^{(k)}} \left|(Q-Q_n)\min_{i\in[\![1,k]\!]} l_{t_i}(u)\mathbb{1}_{\|u\|\leq z}\right| &\leq \frac{2}{n}\mathbb{E}_{X_{1:n}}\mathbb{E}_\sigma\left[\sup_{t\in\mathbb{R}_d^{(k)}}\sum_{i=1}^n \sigma_i \min_{j\in[\![1,k]\!]} l_{t_j}(X_i)\mathbb{1}_{\|X_i\|\leq z}\right] \\
&\leq \frac{2R}{n}\mathbb{E}_{X_{1:n}}\mathbb{E}_\sigma\left[\sup_{t\in\mathbb{R}_d^{(k)}}\sum_{i=1}^n \sigma_i \min_{j\in[\![1,k]\!]} \frac{l_{t_j}(X_i)\mathbb{1}_{\|X_i\|\leq z}}{R}\right],
\end{aligned}$$

where the σ_i's are i.i.d. Rademacher variables, independent of the X_i's, and $\mathbb{E}_Y$ denotes expectation with respect to the random variable Y. As in Section A.6, denote, for any subset of functions G, $\mathcal{N}(G,\varepsilon)$ the ε-covering number of G with respect to the metric $L_2(P_n)$. Denote by $\mathcal{F}_k$ the set of functions $x \mapsto \min_{j\in[\![1,k]\!]} \frac{l_{t_j}(x)\mathbb{1}_{\|x\|\leq z}}{R}$, and by $\mathcal{F}$ the set of functions $x \mapsto \frac{l_t(x)\mathbb{1}_{\|x\|\leq z}}{R}$, $t \in \bar{\mathbb{R}}^d$. Since the $x \mapsto \frac{l_{t_j}(x)\mathbb{1}_{\|x\|\leq z}}{R}$ are bounded by 1, we may write, for any $\varepsilon > 0$,

$$\mathcal{N}(\mathcal{F}_k,\varepsilon) \leq \mathcal{N}(\mathcal{F},\varepsilon)^k,$$

as well as

$$\begin{aligned}
\mathcal{N}(\mathcal{F},\varepsilon) &\leq \mathcal{N}\left(\left\{x \mapsto \frac{-2\langle x, m(Q_{t,h})\rangle\,\mathbb{1}_{\|x\|\leq z}}{R}\right\}, \varepsilon/2\sqrt{2}\right) \\
&\qquad \times \mathcal{N}\left(\left\{x \mapsto \frac{(\|m(Q_{t,h})\|^2 + v(Q_{t,h}))\mathbb{1}_{\|x\|\leq z}}{R}\right\}, \varepsilon/2\sqrt{2}\right) \\
&\leq \mathcal{N}(\mathcal{G}_1, \varepsilon/2\sqrt{2}) \times \mathcal{N}(\mathcal{G}_2, \varepsilon/2\sqrt{2}).
\end{aligned}$$

Using [11, Theorem 1] yields

$$\begin{aligned}
\mathcal{N}(\mathcal{G}_1,\varepsilon)) &\leq \left(\frac{2}{\varepsilon}\right)^{2(d+1)C} \\
\mathcal{N}(\mathcal{G}_2,\varepsilon) &\leq \left(\frac{2}{\varepsilon}\right)^{2C}.
\end{aligned}$$

We then deduce

$$\mathcal{N}\left(\mathcal{F}_k, \varepsilon\right) \leq \left(\frac{4\sqrt{2}}{\varepsilon}\right)^{2k(d+2)C}.$$

Using [5, Corollary 13.2] leads to

$$\mathbb{E}_\sigma\left[\sup_{t\in\mathbb{R}_d^{(k)}} \sum_{i=1}^n \sigma_i \min_{j\in[\![1,k]\!]} \frac{l_{t_j}(X_i)\mathbb{1}_{\|X_i\|\leq z}}{R}\right] \leq C\sqrt{k(d+2)n}.$$

Combining these bounds gives the result of Lemma 28. ■

C.4 Proof of Lemma 29

Proof
For $t \in \mathbb{R}_d^{(k)}$, we get that:

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

Let $t \in \mathbb{R}^d$, and denote by $r = \delta_{Q,h}(t)$, $r_n = \delta_{Q_n,h}(t)$, and $z = 2V\sqrt{(p+1)\log(n)}$. We may write

$$\begin{aligned}
\|m(Q_{t,h}) - m(Q_{n\,t,h})\| &\leq \frac{1}{h}\left(\|Qu\mathbb{1}_{\mathrm{B}(t,r)}(u) - Qu\mathbb{1}_{\mathrm{B}(t,r_n)}(u)\| + \|Qu\mathbb{1}_{\mathrm{B}(t,r_n)}(u) - Q_n u\mathbb{1}_{\mathrm{B}(t,r_n)}(u)\|\right) \\
&\leq \frac{1}{h}\left(\|(Q-Q_n)u\mathbb{1}_{\mathrm{B}(t,r_n)}(u)\| + Q\|u\||\mathbb{1}_{\mathrm{B}(t,r_n)} - \mathbb{1}_{\mathrm{B}(t,r)}|(u)\right) \\
&\leq \frac{1}{h}\left(\|(Q-Q_n)u\mathbb{1}_{\mathrm{B}(t,r_n)}(u)\| + zQ|\mathbb{1}_{\mathrm{B}(t,r_n)} - \mathbb{1}_{\mathrm{B}(t,r)}|(u) + Q\|u\|\mathbb{1}_{\|u\|>z}\right).
\end{aligned}$$

Moreover, $Q\left|\mathbb{1}_{\mathrm{B}(t,r)} - \mathbb{1}_{\mathrm{B}(t,r_n)}\right|(u) = |h - Q(\mathrm{B}(t,r_n))| = |Q_n(\mathrm{B}(t,r_n)) - Q(\mathrm{B}(t,r_n))|$.
On the event described in Lemma 25, we have that

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

Thus,

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

As well,

$$\begin{aligned}
\sup_{t\in\mathbb{R}^d} |M(Q_{t,h}) - M(Q_{n\,t,h})| &\leq \frac{1}{h}\left[\left|(Q-Q_n)\|u\|^2\mathbb{1}_{\mathrm{B}(t,r_n)}\right| + Q\|u\|^2|\mathbb{1}_{\mathrm{B}(t,r_n)} - \mathbb{1}_{\mathrm{B}(t,r}|\right] \\
&\leq \frac{1}{h}\left[\left|(Q-Q_n)\|u\|^2\mathbb{1}_{\mathrm{B}(t,r_n)}\right| + Q\|u\|^2\mathbb{1}_{\|u\|>z} + z^2|(Q-Q_n)\mathbb{1}_{B(t,r_n)}|\right].
\end{aligned}$$

Using Lemma 25 again, we get

$$\left|(Q-Q_n)\|u\|^2\mathbb{1}_{\mathrm{B}(t,r_n)}\right| \leq CV^2\sqrt{d}\frac{(p+1)\log(n)^{\frac{3}{2}}}{\sqrt{n}}$$
$$|(Q-Q_n)\mathbb{1}_{B(t,r_n)}| \leq C\sqrt{d}\frac{\sqrt{(p+1)\log(n)}}{\sqrt{n}}$$
$$Q\|u\|^2\mathbb{1}_{\|u\|>z} \leq 2V^2n^{-(p+1)}.$$

Collecting all pieces leads to

$$|\gamma(t,x)-\hat{\gamma}(t,x)| \leq C\|x\|\frac{V(p+1)\log(n)}{h\sqrt{n}} + CV^2\frac{(p+1)^{\frac{3}{2}}\log(n)^{\frac{3}{2}}}{h\sqrt{n}}. \tag{11}$$

At last, since

$$\mathbb{P}\left\{\max_i \|X_i\| \geq z\right\} \leq ne^{-\frac{t^2}{2V^2}} \leq n^{-2p+1},$$

we deduce that

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

■