Distance functions estimation: contributions to support inference and clustering.

Clément Levrard

Université Paris Cité

Mémoire d'habilitation à diriger des recherches

Distance functions estimation: contributions to support inference and clustering

Auteur :
Clément Levrard

Rapporteurs :

Prof. Benoît Cadre	- Université Rennes 2
Prof. Larry Wasserman	- Carnegie Mellon University

Soutenance prévue le 5 Décembre 2022, devant le jury composé de :

Prof. Mikhaïl Belkin	- University of California San Diego
Prof. Gérard Biau	- Sorbonne Université
Prof. Stéphane Boucheron	- Université Paris Cité
Prof. Benoît Cadre	- Université Rennes 2
Prof. Elisabeth Gassiat	- Université Paris Saclay
Prof. Pascal Massart	- Université Paris Saclay
Prof. Alexandre Tsybakov	- ENSAE Paris
Prof. Larry Wasserman	- Carnegie Mellon University

Remerciements

Tout d'abord, je souhaite remercier Benoît Cadre, Larry Wasserman et Stéphane Boucheron, qui ont accepté de rapporter ce manuscrit. Larry, I am very honored that you accepted to review this habilitation thesis. Ensuite, je tiens à remercier aussi Misha Belkin, Gérard Biau, Elisabeth Gassiat, Pascal Massart et Alexandre Tsybakov, qui me font l'honneur de participer au jury de soutenance.

Les travaux présentés dans ce mémoire sont le fruit de collaborations avec plusieurs personnes au sein de différentes équipes. Par ordre chronologique, je tiens à remercier l'équipe Geometrica de Inria Saclay, qui m'a initié à ces problèmes d'inférence topologique et géométrique, et au sein de laquelle j'ai trouvé des co-auteurs privilégiés (Eddie, Claire, Martin et Fred) ainsi que beaucoup d'idées intéressantes issues de discussions avec Bertrand, Théo, Marc, Steve, Matthieu, Thomas et bien d'autres. Ces recherches se sont poursuivies à Paris 7, où j'ai eu la chance de trouver d'autres interlocuteurs sur ces thèmes, en la présence d'Aurélie bien sûr mais aussi de Sylvain, Mathilde, Ilaria. En dehors de la recherche, c'est encore à Paris 7 que j'ai pu me former au métier particulier d'enseignant ainsi qu'à la mécanique administrative du monde universitaire. J'ai pu y compter sur les compétences et la bienveillance de tout un laboratoire, Stéphane, Francis, Mathieu et Dominique en particulier. Notre petite communauté scientifique s'élargissant au fil des ans, j'ai croisé le chemin de plusieurs personnes dont je suis redevable à bien des égards. En premier lieu Catherine, mais aussi Clément, Marc, Gilles, Olympio, Magalie, et j'en oublie beaucoup qui j'espère voudront bien m'excuser. Enfin, je veux adresser ici des remerciements particuliers à Gérard et Pascal: me trouvant maintenant de l'autre côté de la barrière, je peux mesurer le risque du pari qu'ils ont fait en acceptant de me prendre en thèse.

Les dernières lignes de ces remerciements sont pour la bande d'autodésignés blaireaux, ma famille normande élargie, Lucie et Gabriel. Décrire ce que je leur dois prendrait plusieurs années de psychanalyse.

Contents

Foreword

This report summarizes most of the work that I carried out since my PhD (2011-2014). Co-advised by G. Biau and P. Massart, the central theme of this PhD was theoretical aspects of high-dimensional clustering problems. Technically speaking, these three years complete my former "classical" formation in Statistics via a thorough study of M-estimation problems and related empirical processes techniques.
The year (2014-2015) I spent in the Geometrica (now DataShape) team at Inria Saclay as a post-doctoral fellow has made me discover a rich field of interfaces between Statistics and Geometry/Topology. In addition to specific estimation problems such interfaces naturally carry (estimation of manifolds, curvature, persistence diagrams, etc.), dedicated methods in computational geometry may also provide original approaches to classical problems in Statistics. To me, the benefits of such an interplay are most valuable, so that a significant part of my research work from 2014 until now is focused on these themes.
At last, I was hired in 2015 as Assistant Professor in Université Paris Diderot (now Université de Paris), that is my current position. Interaction with the local research team (LPSM) lead to several lines of research work mostly dealing with quantization and robustness, that are also related to geometric inference problems mentioned in the above paragraph.
The different themes are not equally developed, in particular I chose not to detail the theoretical aspects of k-means clustering that are straight continuation of my PhD. The interested reader could find these details in [Levrard, 2014, 2018, Chazal et al., 2021].

Summary

Most of my recent work pertains to geometric inference or quantization. Distance inference may be thought of as a common thread between these two domains, that allows to understand the difference of targets, proximity measures and means they induce. Though none of these fields can be reduced to distance estimation, we chose to emphasize this particular aspect and the corresponding results.

Chapter 1: Support Estimation

Chapter 1 is devoted to set estimation: the target is d_K (the distance function to K), for a compact set K embedded in $\mathbb{R}^D$ and sampled by $\{X_1, \dots, X_n\} = \mathbb{X}_n$. This problem is one of the salient interfaces between Geometry and Statistics mentioned above.
From a Computational Geometry point of view, some dedicated fields such as curve and surface reconstruction [Dey, 2007] aim at building meshes from a deterministic point cloud that are close enough to K in terms of Hausdorff distance (that is, for a mesh $\hat{K}$ built from a deterministic sample $\mathbb{X}_n$, controlling the term $\|\mathrm{d}_{\hat{K}} - \mathrm{d}_K\|_\infty$). In turn, practical applications of surface meshing can be found in medical imaging, computer graphics and engineering among others. An important feature of these meshes is that their topological structure may be computationally investigated and proved close to that

of K in some particular instances (see [Boissonnat et al., 2018] for a comprehensive introduction).
From a Statistics viewpoint, the sample $\mathbb{X}_n$ is random, with common distribution P that has support K. This support inference problem may be declined into various statistical applications, depending on the structure of K. For instance, seeking connected components of K has clear connection with clustering [Cuevas et al., 2001], whereas retrieving the boundary of a 2-dimensional $K \subset \mathbb{R}^2$ has applications in image analysis [Solé et al., 2004]. A special interest has been recently paid to the case where K has a much smaller dimension d than the ambient space. In this case, estimation of K may be thought of as a non-linear dimensionality reduction technique. This intuition gave birth to numerous dimensionality reduction algorithms via embedding, such as Isomap [Tenenbaum et al., 2000] and LLE Roweis and Saul [2000].
Chapter 1 consists in a panoramic view on minimax rates for support estimation under various regularity assumptions, that provides a frame to our results in the manifold case (where K is a d-dimensional manifold). This still active line of work has been carried out in collaboration with E. Aamari and C. Aaron.
To start with, we introduce the base convergence rate $n^{-1/d}$ that is minimax over the set of distributions that have d-dimensional supports, and is achieved by the naive estimator $\hat{K} = \mathbb{X}_n$, that is the sample itself [Devroye and Wise, 1980]. Then, we expose several convexity-type assumptions on the support that can guarantee improved convergence rates in the full-dimensional case. Namely, in the case where K is D-dimensional and satisfies a convexity-type assumption, then some compact estimators $\hat{K}$ can be built that satisfy

$$\mathbb{E}(\mathrm{d_H}(\hat{K}, K)) \lesssim n^{-\frac{2}{D+1}},$$

provided that the boundary ∂K is smooth enough. Importantly, the proposed estimators rely on constructive and geometrically flavored procedures based on sample points that can be easily implemented. From the theoretical side, the key point is that both regularity of the compact K and its boundary ∂K are needed to ensure that better estimators than $\mathbb{X}_n$ can be conceived.
At last, we expose our work in the general manifold case, when K is assumed to be a d-dimensional manifold, for $d \leq D$. Based on the intuition provided by the full-dimensional case, key regularity parameters of the targeted manifolds are its reach (Federer [1959]), that generalizes the aforementioned convexity-type conditions, and the reach of its boundary, if any. Our results can be summarized in a few lines: in the particular case where K is a compact $\mathcal{C}^k$ manifold without boundary, then some manifold estimators $\hat{K}$ achieve

$$\mathbb{E}(\mathrm{d_H}(\hat{K}, K)) \lesssim n^{-\frac{k}{d}},$$

where the ambient dimension D plays no role (even in the constants). This rate is also proved minimax optimal, and in the case where $k = 2$, a constructive estimator $\hat{K}$ based on the Tangential Delaunay Complex [Boissonnat and Ghosh, 2014] is given that is also topologically consistent. This is also the case for the estimator given in the recent work [Divol, 2021], based on the Čech complex of $\mathbb{X}_n$.
In the case where K is a manifold with non-empty boundary, we provide convergence rates for the $\mathcal{C}^2$ case, showing that

$$\mathbb{E}(\mathrm{d_H}(\hat{K}, K)) \lesssim n^{-\frac{2}{d+1}},$$

for some estimator $\hat{K}$ that relies on local linear patches and a preliminary boundary detection step. This rate is in line with the convergence rates for the full-dimensional

case, and is also proved minimax optimal. The general $\mathcal{C}^k$ case with boundary remains an open question, our current work suggests that a rate in $n^{-\frac{k}{k+d-1}}$ is likely to be optimal.
We conclude this chapter with an overview of results in noisy cases, that is where sample points $\mathbb{X}_n$ are not necessarily drawn on K. Three noise cases are exposed: ambient noise (a portion β of the sample is drawn on K, the remaining part is drawn from a uniform distribution in the ambient space), bounded tubular noise (a sample point X_i takes the form $Y_i + \varepsilon_i$, where Y_i is on K and ε_i is orthogonal to K), and general additive noise. While the two first cases still yields polynomial decay of convergence rates ($n^{-\frac{2}{d}}$ for the ambient noise model for $\mathcal{C}^2$ manifolds, $n^{-\frac{2}{d+2}}$ for the tubular noise model), the general additive noise model presents a logarithmic decay of the minimax rate (at most $\log(n)^{-1}$ if the noise is Gaussian) that mitigates the interest of support estimation as a preliminary step for further inference in this case.

Chapter 2: Coresets and Quantization for distance to support estimation

Chapter 2 motivates the introduction of the notion of coreset for distance to compact functions, and exposes how quantization theory can be adapted to provide such coresets in noisy situations, based on [Brécheteau and Levrard, 2020].
The notion of coreset is tied to a branch of the computational geometry field, that intends to build ε-approximations of various quantities such as volumes, diameters, etc. based on the least possible inputs. Following [Agarwal et al., 2005], a general definition of this notion could be "the extraction of small amount of most relevant information from the given data, and performing the computation on this extracted data".
In the case where the targeted quantity is the distance to compact set d_K, a coreset will be a set of points $\mathbf{c} = (c_1, \ldots, c_k)$ and weights $\omega = (\omega_1, \ldots, \omega_k)$ such that $\mathrm{d}_{\mathbf{c},\omega} : x \mapsto \min_{j=1,\ldots,k} \sqrt{\|x - c_j\|^2 + \omega_j^2}$ is a good approximation of d_K. This particular form has the advantage of allowing for topological inference from the sub-level sets of $\mathrm{d}_{\mathbf{c},\omega}$ [Chazal et al., 2011], as well as decreasing the computational cost of algorithms that construct topological descriptors. For instance, if the topological descriptors are retrieved from a Rips complex filtration of n points (that requires computation of $O(n^2)$ pairwise distances), then the (weighted) Rips filtration based on $\mathrm{d}_{\mathbf{c},\omega}$ would rather require $O(k^2)$ computations.
This chapter begins with a brief presentation of the topological descriptors that motivates the introduction of our particular form of coresets. Some of them, such as Persistence Diagrams, are now common tools from the Topological Data Analysis field (TDA) that have been proved to be relevant features in many statistical learning applications [Singh et al., 2008, Carrière et al., 2019]. Then, preliminary results are given for the conception of coresets from sample in the noise-free case for the sup-norm, that are consequences of Chapter 1: for a precision parameter ε and a sample $\mathbb{X}_n$ drawn on a d-dimensional K, a coreset of size $k(\varepsilon) = O(\varepsilon^{-\frac{1}{d}})$ may be extracted from $\mathbb{X}_n$ using the Farthest Point Sampling Algorithm (choose a first point c_1 at random from $\mathbb{X}_n$, then pick c_2 as the furthest sample point from c_1, c_3 as the furthest sample point from $\{c_1, c_2\}$, etc., and stop when $\mathrm{d}(c_{k+1}, \{c_1, \ldots, c_k\})$ is smaller than ε). A standard volume argument ensures that $k(\varepsilon)$ is optimal, in the sense that any coreset $\mathrm{d}_{\hat{\mathbf{c}},\hat{\omega}}$ built from $\mathbb{X}_n$ that satisfy $\mathbb{E}\|\mathrm{d}_{\hat{\mathbf{c}},\hat{\omega}} - \mathrm{d}_K\|_\infty \leq \varepsilon$ must have $\Omega(\varepsilon^{-\frac{1}{d}})$ points.
In additive noise settings, choosing as coreset a subsample of $\mathbb{X}_n$ is no more relevant. Chapter 2 then follows by introducing quantization as a way to build coresets in this noisy case. To be more precise, quantization aims at designing so-called codebooks $\mathbf{c} = (c_1, \ldots, c_k)$ that minimize $\mathbf{c} \mapsto P(du)V(\mathrm{d}_{\mathbf{c}}(u))$ rather than $\|\mathrm{d}_{\mathbf{c}} - \mathrm{d}_{\mathrm{Supp}(P)}\|_\infty$, where $P(du)f(u)$ denotes integration of f with respect to P and V is a non-decreasing

function. This general principle originates from signal compression theory [Gersho and Gray, 1991], and finds its natural application in source coding [Leis, 2019] and numerical integration [Pagès, 1998]. Usual choices for the potential V are power functions, whose exponent is closely tied with the specification of noise (exponent r is suited for signal perturbation seized in r-Wasserstein distance), this chapter focus on the special case $r = 2$, that intends to minimize $R_P(\mathbf{c}) = P(du)\mathrm{d}^2_{\mathbf{c}}(u)$. This squared-distance choice is motivated by the following facts. First, [Clarkson, 2006] ensures that optimal codebooks in the sense of R_P are optimal coresets for the sup-norm, in the noise-free case. Second, it allows for perturbations seized in W_2 distance, that is $|\sqrt{R_P(\mathbf{c})} - \sqrt{R_Q(\mathbf{c})}| \leq W_2(P, Q)$, so that minimizers of a perturbed version R_Q are $O(W_2(P, Q))$-close to noise-free minimizers, in terms of R_P. At last, codebooks may be built from sample using fast and well-known algorithms such as [Lloyd, 1982].
Unfortunately, optimal codebooks in the sense of R_Q can be proved arbitrarily bad coresets for $\mathrm{d}_{\mathrm{Supp}(P)}$ in some cases, that is $\|\mathrm{d}_{\mathbf{c}^*_Q} - \mathrm{d}_{\mathrm{Supp}(P)}\|_\infty \to +\infty$, where $\mathbf{c}^*_Q$ minimizes R_Q and $k \to +\infty$. Intuitively, this is due to the fact that $P \mapsto \mathrm{d}_{\mathrm{Supp}(P)}$ is not continuous with respect to Wasserstein perturbation.
To address the instability issue of $\mathrm{d}_{\mathrm{Supp}(P)}$, as well as to bypass the $\log(n)^{-1}$ convergence rate for estimating it under Wasserstein perturbation, a standard approach is to introduce a surrogate $\tilde{\mathrm{d}}$ to $\mathrm{d}_{\mathrm{Supp}(P)}$ that is close enough to convey topological guarantees and that can be estimated from sample with faster rates. Such an approach has been proved fruitful for instance in Genovese et al. [2014], Fefferman et al. [2018], where surrogates for the distance to a manifold are proposed that can be estimated at a polynomial rate. In our Wasserstein perturbation framework, a relevant surrogate choice is the Distance to Measure (DTM) Chazal et al. [2011], that is provably stable with respect to W_2 metric and whose bias term may be controlled in the models exposed in Chapter 1, that is $\|\mathrm{d}_{P,h} - \mathrm{d}_{\mathrm{Supp}(P)}\|_\infty \to 0$ whenever $h \to 0$, where $\mathrm{d}_{P,h}$ denotes the DTM. The remaining of Chapter 2 exposes joint work with C. Brécheteau that builds a coreset for the DTM, based on quantization.
Our coreset for the DTM relies on the fact that the distance to measure may be expressed as a power distance, that is

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

The k-Power Distance to Measure (k-PDTM) we propose as a coreset is straightforwardly $\mathrm{d}_{P,h,k}(x) = \min_{j=1,\dots,k} \|x - c^*_j\|^2 + \omega^2_{P,h}(c^*_j)$, where $\mathbf{c}^*$ minimizes the least-square criterion $\mathbf{c} \mapsto P(du)[\min_{j=1,\dots,k} \|u - c_j\|^2 + \omega^2_{P,h}(c_j)]$. This k-PDTM has robustness properties with respect to Wasserstein noise, as for the DTM, and bounds on $\|\mathrm{d}_{\mathrm{Supp}(P)} - \mathrm{d}_{Q,h,k}\|_\infty$ may be stated for the models of Chapter 1 that can exploited for topological inference whenever $W_2(P, Q)$ is small enough. At last, $\mathrm{d}_{Q,h,k}$ may be estimated from sample at a parametric rate in terms of distortion using an empirical risk minimizer, that is $|Q(du)(\mathrm{d}^2_{Q_n,h,k} - \mathrm{d}^2_{Q,h,k})(u)| = O(n^{-\frac{1}{2}})$, that can be used to get a final guarantee on $\|\mathrm{d}_{\mathrm{Supp}(P)} - \mathrm{d}_{Q_n,h,k}\|_\infty$. From a computational viewpoint, $\mathrm{d}_{Q_n,h,k}$ may be approximated using a straightforward adaptation of the Lloyd quantization algorithm [Lloyd, 1982], that is proved efficient in numerical experiments.

Chapter 3: Outlier detection and Clustering

Our last chapter presents two extensions of the coreset designing via quantization principle exposed in Chapter 2. First, we propose an adaptation and generalization of it in the case where adversarial noise is allowed and for dissimilarity measures that can be expressed as Bregman divergences. Second, we depart from the distance function estimation framework to expose a vectorization and compression scheme in the case where data samples come as measures, based on quantization as well.

Robust Bregman Quantization

The first part of Chapter 3 is motivated by the following two remarks. First, the coreset designing problem for the DTM exposed in Chapter 2 can be reduced to minimize $\mathbf{c} \mapsto Q(du)\mathrm{d}_\phi(u, \mathbf{c})$, for a particular Bregman divergence d_ϕ. Second, though the aforementioned k-PDTM adapts to additive noise, it is provably highly sensitive to adversarial perturbations. In collaboration with C. Brécheteau and A. Fischer, we propose a general Robust Bregman Quantization scheme that intends to answer these two remarks. Apart from the robustification of the procedure exposed in Chapter 2, a major interest of this procedure relies in its applications to robust clustering, that is a field of growing interest in the last decades [Banerjee and Davé, 2012, Brunet-Saumard et al., 2022, Cardot et al., 2013, Fritz et al., 2012].

The base idea is simple: for a prescribed threshold α (thought of as the maximal proportion of adversarial noise that can be faced), we intend to minimize the trimmed Bregman Distortion $\mathbf{c} \mapsto R_\alpha(\mathbf{c}) = \min_{A|Q(A)\geq 1-\alpha} Q(du)[\mathrm{d}_\phi(u, \mathbf{c})\mathbb{1}_A(u)]$ from sample, using an empirical risk minimization strategy. In other words, we seek both a coreset $\hat{\mathbf{c}}_{n,\alpha}$ and a trimming set A outside which points are considered as noise, by minimizing $\min_{A|Q_n(A)\geq 1-\alpha} Q_n(du)[\mathrm{d}_\phi(u, \mathbf{c})\mathbb{1}_A(u)]$, where Q_n is the empirical distribution associated with the (possibly corrupted) sample. The robustness of an empirical distortion minimizer $\hat{\mathbf{c}}_{n,\alpha}$ is assessed in two ways: first by proving oracle inequalities of the type

$$R_\alpha(\hat{\mathbf{c}}_{n,\alpha}) - \inf_{\mathbf{c}} R_\alpha(\mathbf{c}) \lesssim n^{-\frac{1}{2}},$$

under an order 2 moment condition that is usual in the robust estimation literature [Brownlees et al., 2015, Catoni and Giulini, 2018]. Second, robustness with respect to adversarial noise is seized in terms of Finite Sample Breakdown Point (FSBP, [Donoho and Huber, 1983]), the minimal portion of adversarial noise that is required to render an estimator arbitrarily bad. In our robust quantization setting, the FSBP of $\hat{\mathbf{c}}_{n,\alpha}$ is closely tied to the structure of Q: indeed, let $p_{j,\alpha}$ is the Q_n-mass of $\{x \in \mathbb{R}^D \mid \forall i \neq j \quad \mathrm{d}_\phi(x, \hat{c}_{n,\alpha,i}) \geq \mathrm{d}_\phi(x, \hat{c}_{n,\alpha,j})\} \cap A(\hat{\mathbf{c}}_{n,\alpha})$, where $A(\hat{\mathbf{c}}_{n,\alpha})$ denotes the optimal trimming set associated with $\hat{\mathbf{c}}_{n,\alpha}$, and $p_{\min,\alpha}$ is the smallest of these Q_n-masses. Then, if $\alpha > p_{\min,\alpha}$, replacing the portion $p_{\min,\alpha}$ of the sample that corresponds to the trimmed Bregman Voronoi cell with the smallest weight by a mass $p_{\min,\alpha}$ located at $\{x_o\}$, where x_o is far enough from the support of Q, will enforce one code point to be located at x_o. Thus the FSBP is smaller than $p_{\min,\alpha}$. On the other hand, it is immediate that the FSBP is smaller than α. We provide a precise bound on the FSBP in terms of structural quantities of Q, that we prove optimal.

In a clustering framework, that is when signal sample is organized around k_0 natural centers, the FSBP may be interpreted as the minimal mass a spurious noise cluster must have to be considered as signal by our procedure. In practice, the choice of a trim level set α and a number k of clusters are interleaved issues: small clusters labeled as signal for small α's may be labeled as noise for larger α's. We propose a screening procedure to identify relevant choices of k (number of clusters) and α from data, that is illustrated on a stylometric author clustering task based on word counting. Interestingly, choosing as Bregman divergence the one associated with the potential $\phi : u \mapsto u\log(u) - u$ (Poisson divergence) yields excellent results for this problem. From a theoretical point of view, the choice of a Bregman divergence may be interpreted as a choice of a distribution from the exponential family, the sample being then modeled as drawings from a mixture of those distributions [Banerjee et al., 2005b]. To some extent, our Robust Bregman Quantization method may be thought of as a robust Classification EM method [Celeux and Govaert, 1992] with arbitrary distributions from the exponential family, in a robust clustering framework.

Quantization-based embedding of measures

We conclude Chapter 3 by presenting an application of quantization to the case where the sample $\mathbb{X}_n$ consists of n i.i.d. *measures* X_i rather than points. The motivating example of this work is TDA-based learning, in which persistence diagrams play the role of features and are thought of as discrete measures on $\mathbb{R}^2$. This situation may also be encountered whenever data are spatial point patterns, as in species distribution modeling [Renner et al., 2015], crime distribution [Shirota et al., 2017], etc., or as bags of instances, this last case being referred to as Multiple Instance Learning (MIL) [Dietterich et al., 1997]. In TDA-based learning, persistence diagrams are often processed via a *vectorization* scheme that intends to convert the (discrete) measures into vectors that can be used in standard learning algorithms. This vectorization step can be carried out by constructing finite-dimensional embeddings [Adams et al., 2017, Carrière et al., 2019] or kernel embeddings [Bubenik, 2015, Carrière et al., 2017]. Most of the more recent vectorization methods are intended to provide the "best" possible vectorization for some dedicated problem, that is often supported by numerical investigations. The approach we propose originates from F. Chazal's empirical remark: "running a k-means on the whole set of points and transforming measures into vector of counts points around centers yields not that bad results". The line of work [Royer et al., 2021, Chazal et al., 2021] we conducted, in collaboration with F. Chazal and M. Royer, backs this intuition. The vectorization process we propose is simple: for a codebook $\mathbf{c} = c_1, \dots, c_k$, a kernel ψ and a bandwidth σ, we embed measures into $\mathbb{R}^k$ via

$$v_i = v(X_i) = (X_i(du)\psi(\|u - c_1\|/\sigma), \dots, X_i(du)\psi(\|u - c_k\|/\sigma)).$$

In practice, the kernel $\psi(u) = \exp(-u)$ gives satisfactory results [Royer et al., 2021], but other choices are possible that convey the same theoretical guarantees [Chazal et al., 2021]. The vectorization codebook is built using quantization algorithms that intend to approximate the mean measure $\mathbb{E}(X)$ (defined by $\mathbb{E}(X)(A) = \mathbb{E}(X(A))$, for a measurable set A, where X is a random measure with the same distribution as X_i, $i = 1, \dots, n$). Intuitively speaking, this vectorization scheme encodes the mass that X_i spreads around every code point c_j. In a mixture model with two components, with distributions $X^{(j)} \sim X_i \mid \{Z_i = j\}$, $j \in \{1, 2\}$, where Z_i is the label, if a code point c falls in an area where $X^{(1)}$ and $X^{(2)}$ differ, then the resulting coordinate on the vectorization will be discriminative. More precise results on discriminative code points, that relate areas of discrimination, kernel and bandwidth choice are stated, that provide guarantees on the vectorization output in a clustering framework, provided one code point fall into such a discriminative area. Of course, similar result can be derived in a supervised learning framework.
Our theoretical targeted codebook is an optimal codebook for $\mathbb{E}(X)$, that is

$$\mathbf{c}^* \in \arg \min_{\mathbf{c} \in (\mathbb{R}^D)^k} \mathbb{E}(X)(du) \min_{j=1,\dots,k} \|u - c_j\|^2.$$

In view of the following learning task, ensuring that optimal codebooks have code points in discriminative areas is crucial. Again, the intuition is simple: in the two-components mixture model considered above, a discriminative area A^* must have positive $X^{(j)}$ mass, for some $j \in \{1, 2\}$ and w.h.p.. Thus, A^* has positive $\mathbb{E}(X)$ mass. Since optimal codebooks are coverings of the support of $\mathbb{E}(X)$ (Clarkson [2006]), an optimal code point must fall in A^* provided k is large enough. This intuition may be quantified, for instance in the case where X is a mixture distribution of persistence diagrams from different shapes. Though the exposed heuristic also works if $\mathbf{c}^*$ is chosen as a uniform grid of the ambient space, let us emphasize that the quantization approach yields the same benefits as for the distance estimation framework of Chapter 2. First it adapts to the intrinsic

dimension (of the support of $\mathbb{E}(X)$), thus requiring less code points than uniform gridding. Second, it may adapt to small additive noise, contrary to choosing code points as a grid of the support of observed measures. In some sense, choosing code points via quantization may be thought of as middle road between uniform grid and sample gridding, the flaws of each being mitigated. This point is illustrated via numerical investigation, showing that our quantization-based approach indeed takes the best of these two gridding methods.
The next step is to build approximations of optimal $\mathbf{c}^*$'s from (measure) sample. To this aim, simple quantization algorithms designed for the point sample [Lloyd, 1982, MacQueen, 1967] case may be adapted to our measure-sample case, resulting in a batch and mini-batch algorithm. From a theoretical point of view, following [Bottou and Bengio, 1995], the aforementioned algorithms may be thought of as a Newton gradient descent (batch version) and a stochastic gradient descent for the risk function $R : \mathbf{c} \mapsto \mathbb{E}(X)(du) \min_{j=1,\dots,k} \|u - c_j\|^2$, so that standard conditions such as co-coercitivity and Lispchitz gradients may ensure convergence of those algorithms. Interestingly, a straightforward adaptation of the margin condition [Levrard, 2015] to the measure sample case is sufficient to guarantee

$$\left\langle D(\mathbf{c})^{-1}\nabla_{\mathbf{c}}R, \mathbf{c} - \mathbf{c}^* \right\rangle \geq (1 - \varepsilon_0)\|\mathbf{c} - \mathbf{c}^*\|^2 \quad \text{(co-coercitivity)},$$
$$\left\| D(\mathbf{c})^{-1}\nabla_{\mathbf{c}}R \right\| \leq (1 + \varepsilon_0)\|\mathbf{c} - \mathbf{c}^*\| \quad \text{(Lipschitz Gradient)},$$

where $D(\mathbf{c})^{-1}$ is a suitable (almost Newton) renormalization used in the algorithms and ε_0 is a small enough constant, provided that $\mathbf{c}$ (thought of as a starting point of the algorithm) is close enough to an optimal $\mathbf{c}^*$. Thus, provided that $\mathbb{E}(X)$ satisfies a margin condition and that initialization is close enough to an optimal $\mathbf{c}^*$, the output of the batch algorithm (stopped after $2\log(n)$ iterations) and the minibatch algorithm (single-pass) both satisfy

$$\|\hat{\mathbf{c}}_n - \mathbf{c}^*\| \lesssim n^{-\frac{1}{2}},$$

that is the minimax rate over the class of distributions satisfying margin conditions [Levrard, 2018].
From an empirical point of view, complexities of the aforementioned quantization algorithms ($O(n\log(n))$ for the batch one and $O(n)$ for the single-pass one) suggest a practical interest in large-scale measure classification or clustering problems. This claim is backed by extensive numerical experiments, including large-scale graph classification for which our global scheme (quantization $\to$ vectorization $\to$ classification) shows close to the state of the art performance, with a much milder computational cost than dedicated methods. To summarize, the (fully unsupervised) vectorization we propose is simple enough to adapt a broad range of situations and scales, at the price of some (experimentally reasonable) loss of optimality. Investigating supervised counterparts of this simple vectorization scheme is ongoing work with O. Hacquard and G. Blanchard.

Publications list

Published

- Oracle inequalities for computationally budgeted model selection, A. Agarwal, P. Bartlett, J. C. Duchi and C. Levrard, *COLT 2011*,
- Fast rates for empirical vector quantization, C. Levrard, *Electronic Journal of Statistics*, 2013,

- Non-asymptotic bounds for vector quantization in Hilbert spaces, C. Levrard, *The Annals of Statistics*, 2015,
- Sparse Oracle Inequalities for Variable Selection via Regularized Quantization, C. Levrard, *Bernoulli Journal*, 2018,
- Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction, E. Aamari and C. Levrard, *Discrete and Computational Geometry*, 2018,
- Non asymptotic bounds for manifold, tangent spaces, and curvature estimation, E. Aamari and C. Levrard, *The Annals of Statistics*, 2018,
- Quantization/Clustering: when and why does k-means work., C. Levrard, *Journal de la Société Française de Statistiques*, 2018,
- A k-points-based distance for robust geometric inference, C. Brécheteau and C. Levrard, *Bernoulli*, 2019,
- Robust Bregman Clustering, C. Brécheteau, A. Fischer and C.Levrard, *The Annals of Statistics*, 2020,
- ATOL: Measure Vectorization for Automatic Topologically-Oriented Learning, M. Royer, F. Chazal, C. Levrard, Y. Umeda, Yuhei and Y. Ike, *Proceedings of Machine Learning Research*, 2020,
- Clustering of measures via mean measure quantization, F. Chazal, C. Levrard and M. Royer, *Electronic Journal of Statistics*, 2021,
- Topologically penalized regression on manifolds, O. Hacquard, K. Balasubramanian, G. Blanchard, C. Levrard and W. Polonik, *Journal of Machine Learning Research*, 2022.

Unpublished

- Minimax Boundary Estimation and Estimation with Boundary, E. Aamari, C. Aaron and C. Levrard, 2021.
- Optimal Reach Estimation and Metric Learning, E. Aamari, C. Berenfeld and C. Levrard, 2022.

Notations

We list here some notation that are used throughout the manuscript.

- $\mathrm{B}(x,r)$ (resp $\bar{\mathrm{B}}(x,r)$), open (resp. closed) ball centered at $x \in \mathbb{R}^D$, with radius r.
- For $A \subset \mathbb{R}^D$, $\mathrm{B}(A,r) = \{x \in \mathbb{R}^D \mid \mathrm{d}(x,A) < r\}$ (resp. $\bar{\mathrm{B}}(A,r)) = \{x \in \mathbb{R}^D \mid \mathrm{d}(x,A) \leq r\}$).
- ∂M, differential boundary, $\bar{\partial} M$ ambient boundary.
- For $K, A \subset \mathbb{R}^D$, $\mathrm{B}_K(A,r) = \mathrm{B}(A,r) \cap K$ (resp. $\bar{\mathrm{B}}_K(A,r) = \bar{\mathrm{B}}(A,r) \cap K)$).
- For $K \subset \mathbb{R}^D$, d_K is the distance to K function.
- For $K, K' \subset \mathbb{R}^D$, $\mathrm{d_H}(K,K')$ is the Hausdorff distance between K and K'.
- For a closed $K \subset \mathbb{R}^D$, π_K is the projection onto K (defined wherever the projection onto K is unique).
- For f a (multi) linear map (resp. matrix, tensor), $\|f\|_{op}$ is the operator norm.
- If T_0, T_1 are two d-dimensional subspaces of $\mathbb{R}^D$, $\angle(T_0,T_1) = \|\pi_{T_0} - \pi_{T_1}\|_{op}$ is the angle between T_0 and T_1.
- For P a measure and f a function, $P(dx)f(x)$ is the integration of f with respect to P (when properly defined).
- For $v \in \mathbb{R}^D$ and $i \in \mathbb{N}^*$, $v^{\otimes i} = (v, \ldots, v) \in \mathbb{R}^{D \times i}$ (tensor product).
- $\mathrm{TV}(P_0, P_1)$, total variation distance between P_0 and P_1 (probability measures).
- $C_{\alpha_1,\ldots,\alpha_p}, c_{\alpha_1,\ldots,\alpha_p}$, large (resp. small) enough constants that only depend on the parameters $\alpha_1, \ldots, \alpha_p$.

Chapter 1

Distance to compact estimation: noise-free case and sup-norm

This chapter contains a non-exhaustive survey of results in compact set estimation theory, often referred to as support estimation in Statistics. Minimax convergence rates in terms of Hausdorff distance, of the type

$$\inf_{\text{estimators } \hat{K}} \sup_{K \in \mathcal{K}} \mathbb{E}\mathrm{d}_{\mathrm{H}}(\hat{K}, K),$$

are investigated under several regularity assumptions on the class of compact sets $\mathcal{K}$. This overview starts from relaxed convexity assumptions in the full dimensional case, under which the naive estimator $\hat{K} = \mathbb{X}_n$, that is the sample itself, is no more optimal. A careful analysis of the influence of these regularity parameters allows to design regularity classes in the case where K has smaller intrinsic dimension d than the ambient dimension D, namely in the manifold case. To summarize, the key regularity assumptions in the manifold case are: differential regularity of the manifold, convexity of the manifold (encoded by its reach), and convexity of its boundary.
Section 1.4 gathers and frames the results in Aamari and Levrard [2018, 2019], Aamari et al. [2021] pertaining to manifold estimation, emphasizing connections with the full-dimensional regularity classes exposed in the first sections. To summarize, provided the reach of manifolds and boundaries (if any) are uniformly lower bounded, the minimax convergence rates in terms of Hausdorff distances are $n^{-\frac{k}{d}}$ in the $\mathcal{C}^k$ case without boundary, and $n^{-\frac{2}{d+1}}$ in the $\mathcal{C}^2$-case with boundary. These rates are attained by constructive manifold and boundary estimators that are described throughout the section.
We conclude this chapter by providing a short overview of existing results in noisy cases, giving some insights on what type of noise leads to a difficult estimation problem.

1.1 The compact set estimation framework

Throughout this chapter, we let P be a probability distribution whose support $K \subset \mathbb{R}^D$ is to be inferred. The support K is formally defined via

$$K = \bigcap_{F \text{ closed s.t. } P(F)=1} F.$$

By definition, K is a closed subset of $\mathbb{R}^D$. We make the additional assumption that K is bounded, thus compact.
Given $\mathbb{X}_n = \{X_1, \dots, X_n\}$, an i.i.d. n-sample drawn from P, an estimator $\hat{K}$ of K is a random subset of $\mathbb{R}^D$. The proximity between K and $\hat{K}$ will be seized in terms of

Hausdorff distance,

$$\mathrm{d_H}(\hat{K}, K) = \inf\left\{r \geq 0 \mid \hat{K} \subset \bar{B}(K, r) \quad \text{and} \quad K \subset \bar{B}(K, r)\right\},$$

where, for any set $A \subset \mathbb{R}^D$, $B(A, r)$ (resp. $\bar{B}(A, r)$) denotes the (closed) r-offset of A, that is $B(A, r) = \bigcup_{x \in A} B(x, r)$ (resp. $\bigcup_{x \in A} \bar{B}(x, r)$). If d_A denotes the distance to A function

$$\mathrm{d}_A : \begin{cases} \mathbb{R}^D & \longrightarrow & \mathbb{R}^+ \\ x & \mapsto & \inf_{x \in A} \|x - A\|, \end{cases}$$

then it is easy to see that

$$\mathrm{d_H}(\hat{K}, K) = \|\mathrm{d}_{\hat{K}} - \mathrm{d}_K\|_\infty.$$

This basic equality exposes the equivalence between compact set estimation in terms of Hausdorff distance and distance estimation in terms of sup-norm. To bypass potential measurability issues in the space of compact sets, we assume that

$$\begin{cases} (\mathbb{R}^D)^n, \mathcal{B}\left((\mathbb{R}^D)^n\right) & \longrightarrow & (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\ (x_1, \ldots, x_n) & \mapsto & \mathrm{d_H}(\hat{K}(x_1, \ldots, x_n), K) \end{cases}$$

is measurable.

1.2 The (a, b)-standard assumption

Without additional regularity assumptions on K (and P), it falls under the intuition that the compact set estimation problem is hard in the minimax sense, that is

$$\inf_{\hat{K}} \sup_{\mathrm{Supp}(P) = K \subset B(0,R)} \mathbb{E}\mathrm{d_H}(\hat{K}, K),$$

for some maximal radius R. Indeed, consider $P_0 = \delta_0$ with support $K_0 = \{0\}$, and $P_1 = (1-u)\delta_0 + u\delta_{x_0}$ with support $K_1 = \{0, x_0\}$, where $u \in]0, 1[$. To make the discrimination between K_0 and K_1 based on $\mathbb{X}_n$, a sample point must fall on $\{x_0\}$. Such an event has probability smaller than $1 - (1-u)^n$, therefore, with probability larger than $(1-u)^n$, deciding whether the support is K_0 or K_1 is a random guess. The cost incurred for such a random guess is $\mathrm{d_H}(K_0, K_1)/2 = R/2$. Choosing u small enough ($1/n$ for instance) allows to lower bound the minimax risk. This intuition can be formalized using Le Cam's lemma (see, e.g., Yu [1997]), adapted to our framework.

Lemma 1.1 : [Aamari and Levrard, 2019, Theorem 8]

Let $\mathcal{P}$ be a set of compactly supported distributions. Then, for any $P_0, P_1 \in \mathcal{P}$, with supports K_0 and K_1, we have

$$\inf_{\hat{K}} \sup_{P \in \mathcal{P}} \mathbb{E}\left(\mathrm{d_H}(\hat{K}, K)\right) \geq \frac{1}{2}\mathrm{d_H}(K_0, K_1)(1 - TV(P_0, P_1))^n,$$

where TV *denotes the total variation distance, that is*

$$\mathrm{TV}(P_0, P_1) = \sup_{A \in \mathcal{B}(\mathbb{R}^D)} |P_0(A) - P_1(A)|.$$

Considering P_0 and P_1 described above, it easily follows that

$$\inf_{\hat{K}} \sup_{\mathrm{Supp}(P)=K\subset B(0,R)} \mathbb{E}\mathrm{d_H}(\hat{K}, K) \geq c_0 R,$$

where c_0 is an absolute constant. From this viewpoint, the constant estimator $\hat{K} = \{0\}$ is minimax optimal over the class of compactly supported distribution (whose support is included in a prescribed ball).
To avoid the potential for spurious areas with arbitrarily small weights, a general approach is to require that P spreads enough mass around every point of its support, leading to the following definition.

Definition 1.2 :

> Let a, b be positive real numbers. A measure P on $\mathbb{R}^D$ satisfies the (a, b)-standard assumption at scale r_0 if
>
> $$\forall x \in \mathrm{Supp}(P) \quad \forall r \leq r_0 \quad P(B(x, r)) \geq a r^b \wedge 1.$$
>
> A measure P is called (a, b)-standard if it satisfies the (a, b)-standard assumption at scale r_0, for some $r_0 > 0$.

Roughly, the (a, b)-standard assumption requires that balls with centers in the support have masses that can be compared with Lebesgue volumes of b-dimensional Euclidean balls (whenever $b \in \mathbb{N}$). Note that a more general definition can be found in Cuevas and Fraiman [1997] for instance, allowing comparison with a general measure μ rather than the b-dimensional Lebesgue measure.
This (a, b)-standard assumption is usual in set estimation theory (Cuevas [2009], Cuevas and Rodríguez-Casal [2004], but also in Topological Data Analysis, from the statistics point of view (Fasy et al. [2014], Chazal et al. [2015a,c, 2016]). It encompasses convexity type assumptions such as the ones described in Section 1.3, but also cases where P is a distribution on a smooth structure (such as a $\mathcal{C}^2$ manifolds), with lower bounded density. As described by the following examples, both smoothness of the support and density assumptions are encoded via the (a, b)-standard assumption.

Example 1.3.

1. Assume that P is supported by M, a $\mathcal{C}^2$ d-dimensional submanifold of $\mathbb{R}^D$. Let $\mathcal{H}_M$ denote the Hausdorff volume measure on M ([Federer, 1969, 2.10.2]), and assume that P has a density f with respect to $\mathcal{H}_M$, lower bounded by $f_{\min}$. Then P is (a, d)-standard, for some $a > 0$ (Aamari et al. [2021]).

2. If K is the Cantor set, $\mathcal{H}_K$ is its Hausdorff measure, and P has a lower-bounded density f with respect to $\mathcal{H}_K$, then P is $(a, \log(2)/\log(3))$-standard.

3. If $K = [0, 1]$, and P has density proportional to $[x(1 - x)]_+$, then P is $(a, 2)$-standard, for some $a > 0$.

4. Assume that $K \subset \mathbb{R}^2$ has an arbitrarily narrow corner (as depicted by Figure 1.1), and that P has a density f with respect to the 2-dimensional Lebesgue measure. Then, for any (a, b) such that $b \leq 2$, P cannot be (a, b)-standard.

Provided that the (a, b)-standard assumption holds, consistent estimators of K may be defined. A basic estimator is $\hat{K} = \mathbb{X}_n$, that estimates K directly via the sample point. The following result gives convergence rates for this naive estimator.

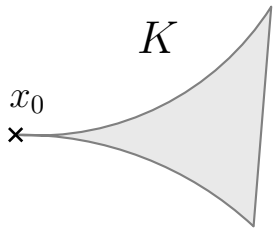


Figure 1.1: Arbitrarily narrow corner at x_0. Uniform distribution on K cannot be $(a, 2)$-standard.

THEOREM 1.4 : [AAMARI, 2017, PROPOSITION III.14]

Assume that P is (a, b)-standard at scale r_0, let K denote the support of P, and $\hat{K} = \mathbb{X}_n$. Denote by

$$r_n = \left(\frac{[(1 + 1/b) \vee 4^b] \log(n)}{an} \right)^{\frac{1}{b}}.$$

If n is large enough so that $r_n \leq 2r_0$, then

$$\mathbb{E}[\mathrm{d_H}(\hat{K}, K)] \leq r_n.$$

Note that deviation bounds can also be found in Chazal et al. [2015d]. Interestingly, the same authors prove that no faster rate of convergence can be expected over the class of (a, b)-standard distributions.

PROPOSITION 1.5 : [CHAZAL ET AL., 2015D, SECTION B.2.2]

Let $\mathcal{P}_{(a,b)-st}$ denote the set of (a, b)-standard probability measures. Then

$$\inf_{\hat{K}} \sup_{P \in \mathcal{P}_{(a,b)-st}} \mathbb{E}\left[\mathrm{d_H}(\hat{K}, K)\right] \geq c_{a,b} \left(\frac{1}{n}\right)^{\frac{1}{b}}.$$

This result is based on Lemma 1.1, with two hypotheses on the real line, namely δ_0 (that is (a, b)-standard for any a and b), and $(1 - 1/n)\delta_0 + (1/n)\delta_{x_n}$, where x_n tends to 0 with rate $n^{-1/b}$. This stresses out that the (a, b)-standard assumption is not completely tied to the geometric structure of the support. Concerning the $\log(n)$ term of the upper bound, we strongly believe that a lower bound in $(\log(n)/n)^{\frac{1}{b}}$ could be derived, at the price of more technicalities as in Kim and Zhou [2015].
Though providing consistent estimators, the (a, b)-standard assumption is still too loose to render the construction of interesting compact estimators relevant, as alleged by Proposition 1.5. Consequently, additional assumptions on the structure of $\mathrm{Supp}(P)$ will be exposed in the following sections, under which the naive estimator $\mathbb{X}_n$ will be no longer minimax optimal.

1.3 Convexity type assumptions

As detailed in the last section, structural assumptions on K are needed to conceive better estimation strategies than the sample points estimator $\mathbb{X}_n$. Several convexity-type properties have been conceived, leading to different estimation strategies. We briefly expose here the notions of r-rolling condition Cuevas et al. [2012], r-convexity Rodríguez Casal [2007b], and bounded reach Federer [1959].

Definition 1.6 :

Let $r > 0$.

- [Cuevas et al., 2012, Definition 1]: K satisfies the r-rolling condition if, for all $x \in \bar{\partial} K$, there exists $p \notin K$ such that $\bar{B}(p, r) \cap K = \{x\}$ and $B(p, r) \cap K = \emptyset$.
- [Rodríguez Casal, 2007b, Definition 1]: K is r-convex ,for $r > 0$, if

$$K = \bigcap_{B(x,r) \subset K^c} B(x, r)^c.$$

Note that the r-convex condition is strictly stronger than the r-rolling ball assumption (see, e.g., [Cuevas et al., 2012, Proposition 2]). To introduce the bounded reach condition, let us first introduce the medial axis

$$\mathrm{Med}(K) = \{x \in \mathbb{R}^D \mid \exists a, b \in K, \quad a \neq b, \quad \|x - a\| = \|x - b\| = d(x, K)\}. \tag{1.1}$$

The reach Federer [1959] of K, τ_K is then defined by

$$\tau_K = \inf_{p \in \mathrm{Med}(K)} d(p, K) = \min_{x \in K} d(x, \mathrm{Med}(K)).$$

Then the bounded reach assumption is $\tau_K > 0$. This assumption is standard in

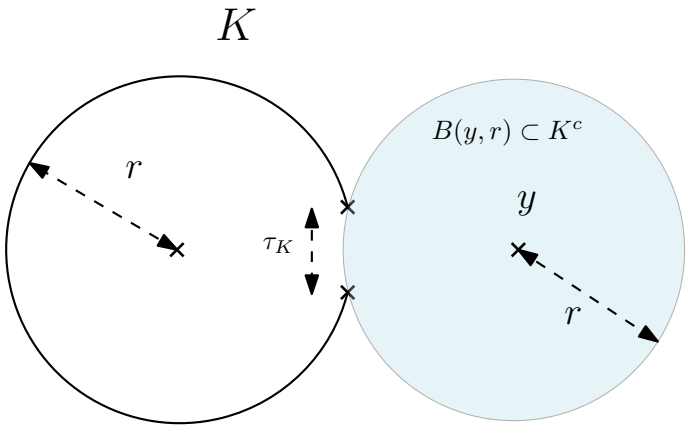


Figure 1.2: r convexity and reach. K is r-convex, but $\tau_K < r$.

topological inference, where τ_K^{-1} is called condition number Niyogi et al. [2008], Singer and Wu [2012]. Note that, for a given $r > 0$, $\tau_K \geq r$ implies the r-convexity assumption, but the converse is not true Cuevas et al. [2012], see Figure 1.2. At last, let us mention that $\tau_K = +\infty$ is equivalent to K is convex [Federer, 1959, Remark 4.2]. For any $r > 0$, implications between these convexity conditions can be summarized as

$$\text{Convexity} \quad \Rightarrow \quad \tau_K \geq r \quad \Rightarrow \quad r\text{-convexity} \quad \Rightarrow \quad r\text{-rolling ball},$$

with strict inclusions. Whenever such a convexity-type assumption is required, more involved estimators can be designed. For instance, if K is convex, then $\hat{K} = Conv(\mathbb{X}_n)$ seems a relevant estimator of K. This intuition may be further pushed in the r-convex case by considering the r-convex hull of the sample point, that is

$$\hat{K} = \bigcap_{\mathrm{B}(x,r) \subset \mathbb{X}_n^c} \mathrm{B}(x, r)^c.$$

However, without further regularity assumptions, these convexity conditions are not enough to provide improved convergence rates. Indeed, let C be a d-dimensional convex

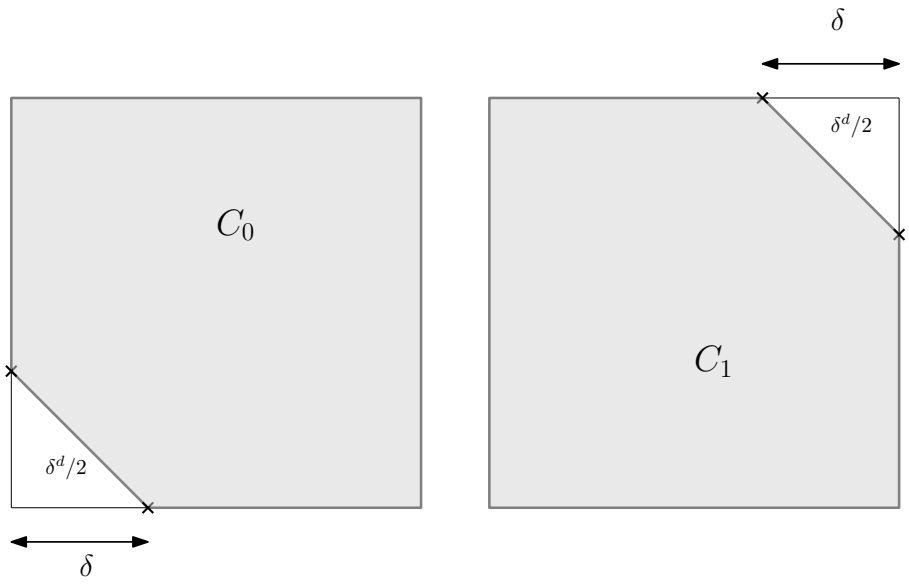


Figure 1.3: Two hypothesis for convex estimation.

set, and assume that P has a density with respect to the d-dimensional Lebesgue measure that is lower bounded by $f_{\min}$. We denote by $\mathcal{P}_{cvx,f_{\min}}$ the set of such probability distributions. Note that a distribution $P \in \mathcal{P}_{cvx,f_{\min}}$ is (a,d)-standard, for some $a > 0$. Intuitively speaking, if C has a corner, then it is likely that taking the convex hull will not yield better results than considering the sample in the corner area. This intuition is formalized by the following result.

Proposition 1.7 :

> *Recall that $\mathcal{P}_{cvx,f_{\min}}$ is the set of distributions with convex support and density $f \geq f_{\min} > 0$. It holds, for $f_{\min}$ small enough,*
>
> $$\inf_{\hat{C}} \sup_{\mathcal{P}_{cvx,f_{\min}}} \mathbb{E}[\mathrm{d}_H(\hat{C}, C)] \geq c_{d,f_{\min}} n^{-\frac{1}{d}}.$$

Such a lower bound has been proved in [Korostelëv and Tsybakov, 1994, Theorem 2] for the special case $d = 2$ (with explicit constants). We provide here a short generalization to the d-dimensional case.

Proof of Proposition 1.7. Following the intuition, we will define two convex sets with corners that will differ at corners (see Figure 1.3). We start from the unit cube $[0,1]^d$. For a fixed $\delta > 0$, we let Λ_δ denote the pyramidal cap defined as the convex hull of $\{0, \delta e_1, \ldots, \delta e_d\}$, where $(e_1, \ldots, e_d)$ denotes the canonical basis of $\mathbb{R}^d$. Then C_0 is defined via $C_0 = [0,1]^d \setminus \Lambda_\delta$, and $C_1 = [0,1]^d \setminus (1 - \Lambda_\delta)$. P_0 and P_1 are defined as the uniform distributions over C_0 and C_1.
It is immediate that $\mathrm{d_H}(C_0, C_1) = \sqrt{d}\delta$. Moreover, for $\delta \leq C_d n^{-\frac{1}{d}}$, we have $TV(P_0, P_1) \leq n^{-1}$. Applying Lemma 1.1 gives the desired result. □

Roughly speaking, Proposition 1.7 shows that no statistical gain can be expected from the additional convexity assumption compared to the (a,d)-standard model. Since the convex assumption is stronger than all aforementioned convexity-type conditions, there is no hope that such conditions could lead to better set estimators than $\mathbb{X}_n$.
To get better convergence rates, we have to ensure that no corner can occur. There exists several assumptions to avoid such a situation, such as the ones given in Dümbgen

and Walther [1996], Bárány [1989], Korostelëv and Tsybakov [1994] for the convex case, all of them requiring that ∂K is smooth, in a certain sense. To ease the intuition, we will assume that ∂K is $\mathcal{C}^2$, that is enough to satisfy the aforementioned conditions. In this case, improved convergence rates can be derived.

THEOREM 1.8 : [DÜMBGEN AND WALTHER, 1996, COROLLARY 1]

Let $P \in \mathcal{P}_{cvx,f_{\min}}$, and let $\hat{K}$ denotes the convex hull of the sample points. It holds

$$\mathrm{d_H}\left(\hat{K}, K\right) = \begin{cases} \mathcal{O}\left((\log(n)/n)^{\frac{1}{d}}\right) & a.s. \\ \mathcal{O}\left((\log(n)/n)^{\frac{2}{d+1}}\right) & a.s. \quad \textit{if } \partial K \textit{ is } \mathcal{C}^2. \end{cases}$$

A more precise result for the case $d = 2$ can be found in [Korostelëv and Tsybakov, 1994, Theorem 3], with explicit constants depending on the curvature of ∂K. Note that the convexity condition can be slightly relaxed: a similar result is given by [Rodríguez Casal, 2007b, Theorem 3] for the r-convex hull in the r-convex case. In these two cases (convex and r-convex), the proposed convergence rates are in fact driven by the estimation rate of the boundary ∂K.
It is important to mention that improved convergence results under regularity assumptions of the boundary hold in the full-dimensional case $d = D$, or in the case where K is included in a d-dimensional linear subspace of $\mathbb{R}^D$, that allows to define the boundary ∂K. The following section extend this flat case, by considering the case of submanifolds of $\mathbb{R}^D$.

1.4 Manifold estimation

Throughout this section we assume that K is a compact d-dimensional submanifold of $\mathbb{R}^D$, that we rename M. In the manifold setting, the boundary may be defined in a differential geometric way. By definition, the d-dimensional submanifolds $M \subset \mathbb{R}^D$ with boundary are the subset of $\mathbb{R}^D$ that can locally be parametrized either by the Euclidean space $\mathbb{R}^d$, or the half-space $\mathbb{R}^{d-1} \times \mathbb{R}_{\geq 0}$ [Lee, 2011, Chapter 2].

DEFINITION 1.9 : SUBMANIFOLD WITH BOUNDARY, BOUNDARY, INTERIOR

A closed subset $M \subset \mathbb{R}^D$ is a d-dimensional submanifold with boundary of $\mathbb{R}^D$, if, for all $p \in M$ and all small enough open neighborhood V_p of p in $\mathbb{R}^D$, there exists an open neighborhood U_0 of 0 in $\mathbb{R}^D$ and a diffeomorphism $\Psi_p : U_0 \to V_p$ with $\Psi_p(0) = p$, such that either:

(i) $\Psi_p\left(U_0 \cap \left(\mathbb{R}^d \times \{0\}^{D-d}\right)\right) = M \cap V_p$. Such a $p \in M$ is called an *interior* point of M, the set of which is denoted by $\mathring{M}$.

(ii) $\Psi_p\left(U_0 \cap \left(\mathbb{R}^{d-1} \times \mathbb{R}_{\geq 0} \times \{0\}^{D-d}\right)\right) = M \cap V_p$. Such a $p \in M$ is called a *boundary* point of M, the set of which is denoted by ∂M.

Remark 1.10. Recall that the *geometric* (or *differential*) boundary ∂M is not to be confounded with the ambient *topological* boundary $\bar{\partial} S := \bar{S} \setminus \mathring{S}$ for $S \subset \mathbb{R}^D$, where the closure and interior are taken with respect to the ambient topology of $\mathbb{R}^D$. One easily checks that if $d < D$, then $\bar{\partial} M = \emptyset$. On the other hand, the two notions $\bar{\partial} M$ and ∂M are guaranteed to coincide when $d = D$. This differential boundary definition naturally

extends the definition of the boundary of a convex set expressed in terms of locations of supporting hyperplanes.

Then, submanifolds *without* boundary are those M that fulfill $\partial M = \emptyset$, i.e. that are everywhere locally parametrized by $\mathbb{R}^d$, and nowhere by $\mathbb{R}^{d-1} \times \mathbb{R}_{\geq 0}$. From this perspective, as confusing as this standard terminology can be, (sub)manifolds without boundary are special cases of (sub)manifolds with boundary. Note that key instances of manifolds without boundary are given by boundaries of manifolds, as expressed by the following result.

PROPOSITION 1.11 : [LEE, 2011, EXAMPLE 2.17]

> *If $M \subset \mathbb{R}^D$ is a d-dimensional submanifold with nonempty boundary ∂M, then ∂M is a $(d-1)$-dimensional submanifold without boundary.*

Concerning smoothness assumptions, we will require that M is $\mathcal{C}^k$, for $k \geq 2$, that is the diffeomorphism of Definition 1.9 are $\mathcal{C}^k$. This allows us to endow M with the Riemannian metric given by the ambient Euclidean metric, that provides a convenient framework for calculations. Further, the reach condition implies results on the structure of this Riemannian metric, as exposed below.

PROPOSITION 1.12 :

> *Let M be a compact $\mathcal{C}^2$ submanifold of $\mathbb{R}^D$. Then the following holds.*
>
> *(i)* $\tau_M > 0$.
>
> *(ii) [Niyogi et al., 2008, Proposition 6.1]: For all p in M, $\|\mathrm{II}\|_{op} \leq \tau_M^{-1}$, where II denotes the second fundamental form of M at point p ([do Carmo, 1992, p.125]).*
>
> *(iii) [Niyogi et al., 2008, Proposition 2.1], [Alexander and Bishop, 2006, Corollary 1.4]: The injectivity radius (see, e.g., [Gallot et al., 2004, Definition 2.116]) of M is at least $\pi\tau_M$.*

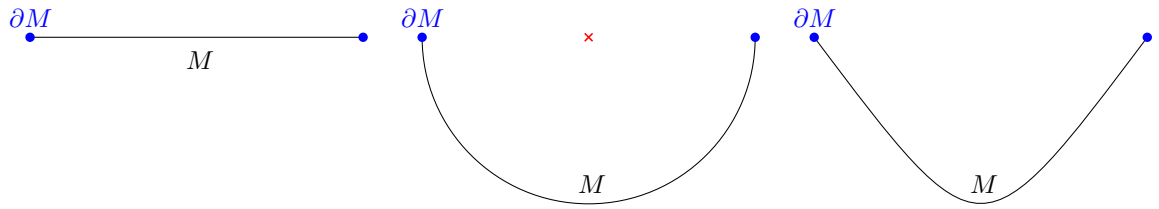
The two last items are well-known results, the first one easily derives from the characterization of the reach given by [Federer, 1959, Theorem 4.18] and basic differential geometry. It follows from our definition of the boundary, Definition 1.9, Propositions 1.11 and 1.12 that ∂M is a $\mathcal{C}^k$ manifold without boundary and positive reach, provided that M is a $\mathcal{C}^k$ manifold with boundary. Thus, for a smoothness order $k \geq 2$, the regularity of M is parametrized by τ_M and $\tau_{\partial M}$. We define our geometric model in the following way.

DEFINITION 1.13 : GEOMETRIC MODEL

> Let $d \leq D$, and k be integers. For positive constants $\tau_{\min}$, $\tau_{\partial,\min}$, we let $\mathcal{M}^{k,d,D}_{\tau_{min},\tau_{\partial,\min}}$ denote the set of $\mathcal{C}^k$ d-dimensional submanifolds $M \subset \mathbb{R}^D$ satisfying
>
> (i) $\tau_M \geq \tau_{\min}$,
>
> (ii) $\tau_{\partial M} \geq \tau_{\partial,\min}$.

Some remarks on the connection of our model to convexity models are well as some insights on the connections between τ_M and $\tau_{\partial M}$ are given below.

Remark 1.14.



(a) $\tau_{\partial M} < \tau_M = \infty$. (b) $\tau_{\partial M} = \tau_M$. (c) $\tau_{\partial M} > \tau_M$.

Figure 1.4: In full generality ($d \leq D$), the reach of a submanifold M and that of its boundary ∂M are not related quantities.

- Let us recall that the model $\mathcal{M}^{k,d,D}_{\tau_{\min},\tau_{\partial,\min}}$ includes both submanifolds with empty and non-empty boundary ∂M, the main requirement on it being that $\tau_{\partial M} \geq \tau_{\partial,\min}$. If $\partial M = \emptyset$, this requirement is always fulfilled since $\tau_\emptyset = \infty$. Note also that Definition 1.13 does not exclude the case $d = D$, in which case M consists of a domain of $\mathbb{R}^D$ with non-empty interior. Furthermore, since the boundary ∂M of a submanifold M is either empty or itself a submanifold without boundary, a non-empty ∂M cannot be convex [Hatcher, 2002, Theorem 3.26]. As a result, $\mathcal{M}^{k,d,D}_{\tau_{min},\infty}$ is exactly the set of submanifolds $M \in \mathcal{M}^{k,d,D}_{\tau_{\min},\tau_{\partial,\min}}$ that have empty boundary. In particular, Definition 1.13 encompasses the model of Genovese et al. [2012b], Kim and Zhou [2015], Aamari and Levrard [2018].

- Similarly, since $\tau_M = \infty$ if and only if M is convex, $\mathcal{M}^{k,d,D}_{\infty,\tau_{\partial,\min}}$ is exactly the set of submanifolds $M \in \mathcal{M}^{k,d,D}_{\tau_{\min},\tau_{\partial,\min}}$ that are convex (and hence have non-empty boundary). In particular, Definition 1.13 encompasses the model of Dümbgen and Walther [1996].

- In full generality, the two lower bounds on the respective reaches of M and ∂M are *not* redundant with one another. As shown in Figure 1.4, τ_M and $\tau_{\partial M}$ can be in any order. However, for $d = D$, ∂M is the topological boundary of M, $\bar{\partial} M$. In this case, [Federer, 1959, Remark 4.2] and an elementary connectedness argument show that $\tau_M \geq \tau_{\partial M}$. Hence, for $d = D$ one may set $\tau_{\min} = \tau_{\partial,\min}$ without loss of generality.

Our geometric model $\mathcal{M}^{k,d,D}_{\tau_{\min},\tau_{\partial,\min}}$ being settled, the sampling scheme on such manifolds may be defined. We recall here that any d-dimensional submanifold of $\mathbb{R}^D$ inherits a natural measure from the d-dimensional Hausdorff measure $\mathcal{H}^d$ on $\mathbb{R}^D$ [Federer, 1969, p.171], that we denote by vol_M.

Definition 1.15 : Statistical Model

> Let $0 < f_{\min} \leq f_{\max}$ be positive numbers. We define $\mathcal{P}^{k,d,D}_{\tau_{\min},\tau_{\partial,\min}}(f_{\min}, f_{\max})$ as the set of Borel probability distributions P on $\mathbb{R}^D$ such that:
>
> - $M = \mathrm{Supp}(P) \in \mathcal{M}^{k,d,D}_{\tau_{\min},\tau_{\partial,\min}}$,
> - P has a density f with respect to the volume measure $\mathrm{vol}_M = \mathbb{1}_M \mathcal{H}^d$ on M, such that $f_{\min} \leq f(x) \leq f_{\max}$ for all $x \in M$.

In a nutshell, we will assume that $\mathbb{X}_n$ is a n-sample drawn almost uniformly on a manifold from $\mathcal{M}^{k,d,D}_{\tau_{\min},\tau_{\partial,\min}}$. This statistical model has clear connections with (a,d)-standard models. Moreover, some of the parameters of this model are in fact related, as detailed below.

(a) P_0. (b) P_1.

Figure 1.5: Two (a, d)-standard hypothesis.

PROPOSITION 1.16 :

Let $P \in \mathcal{P}^{k,d,D}_{\tau_{\min},\tau_{\partial,\min}}(f_{\min}, f_{\max})$ with support M.

(i) [Aamari and Levrard, 2018, Lemma 2] If $\operatorname{diam}(M)$ denotes the diameter of M,

$$\operatorname{diam}(M) \leq \frac{C_d}{\tau_M^{d-1} f_{\min}}.$$

(ii) [Aamari et al., 2019, Proposition 2.7] If M is connected, then

$$\tau_{\min}^d \leq \frac{C_d}{f_{\min}}.$$

(iii) Combining [Aamari et al., 2021, Corollary A.6] and [Aamari and Levrard, 2019, Lemma B.7]: Assume that $0 \leq r \leq \frac{\tau_{\min} \wedge \tau_{\partial,\min}}{40}$. Then, for every $x \in M$,

$$P(B(x, r)) \geq c_d f_{\min} r^d.$$

Though the original results *i)* and *ii)* are stated for the case of manifolds with empty boundary, their proof also hold in the boundary case. The intuition behind these two first points is that, for a prescribed geometric regularity $\tau_{\min}$, the lower bound $f_{\min}$ has to be small enough to have a non-empty model $\mathcal{P}^{k,d,D}_{\tau_{\min},\tau_{\partial,\min}}(f_{\min}, f_{\max})$. Point *iii)* confirms that our model is a submodel of the set of (a, d)-standard distributions, for a small enough a. Note that $\tau_{\min}$ and $\tau_{\partial,\min}$, that encodes respectively smoothness and convexity of the manifold and its boundary, are parameters of key importance to depart from the (a, d)-standard case.

Indeed, consider the two hypothesis depicted by Figure 1.5 in the empty boundary case, and smooth P_1 with a kernel with radius σ, leading to a smoothed peak of height δ_σ and basis δ, where $\delta_\sigma \to 0$ as $\sigma \to 0$. Then, for any $f_{\min}$ small enough and $f_{\max}$ large enough, applying Lemma 1.1 leads to

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}^{k,d,D}_{0,\tau_{\partial,\min}}(f_{\min},f_{\max}) \cap \mathcal{P}_{(a,d)-st}} \mathbb{E}[\mathrm{d_H}(\hat{M}, M)] \geq \frac{1}{2}\delta_\sigma (1 - C_d(\delta + \sigma)^d)^n,$$

for a small enough (depending on $f_{\min}$) and every $\tau_{\partial,\min}$. Choosing $\delta \leq C_d n^{-1/d}$ and σ small enough with respect to n leads to, for a small enough $f_{\min}$, small enough a, and large enough $f_{\max}$,

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}^{k,d,D}_{0,\tau_{\partial,\min}}(f_{\min},f_{\max}) \cap \mathcal{P}_{(a,d)-st}} \mathbb{E}[\mathrm{d_H}(\hat{M}, M)] \geq c_d n^{-1/d},$$

whatever the choice of $\tau_{\partial,\min}$.

Concerning $\tau_{\partial,\min}$, smoothing the hypothesis of the convex case depicted by Figure 1.3, with a smoothing parameter that tends to 0 gives

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}^{k,d,D}_{\tau_{\min},0}(f_{\min},f_{\max}) \cap \mathcal{P}_{(a,d)-st}} \mathbb{E}[\mathrm{d_H}(\hat{M}, M)] \geq c_d n^{-1/d},$$

for $f_{\min}$, a small enough and $f_{\max}$ large enough, whatever the choice of $\tau_{\min}$. This demonstrates that controlling both the regularity of M and ∂M via $\tau_{\min}$ and $\tau_{\partial,\min}$ is a crucial assumption to obtain improved convergence rates. In other words, not controlling τ_M or $\tau_{\partial,M}$ brings us back to the (a,d)-standard model.
Note that if the hypothesis are allowed to go out the (a,d)-standard model (that is not the case in the two aforementioned examples), then adding an arbitrarily narrow smooth peak to a boundaryless manifold or an arbitrarily narrow corner to a manifold with boundary leads to statistical undecidability, that is

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}^{k,d,D}_{0,\tau_{\partial,\min}}(0,f_{\max})} \mathbb{E}[\mathrm{d_H}(\hat{M}, M)] \geq c_d,$$

for $f_{\max}$ large enough and every $\tau_{\partial,\min}$, and

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}^{k,d,D}_{\tau_{\min},0}(0,f_{\max})} \mathbb{E}[\mathrm{d_H}(\hat{M}, M)] \geq c_d,$$

for $f_{\max}$ large enough and $\tau_{\min}$ small enough.
We are now in position to investigates convergence rates for support estimation over these regularity classes. We begin with the particular case of manifolds without boundary.

1.4.1 $\mathcal{C}^2$ manifolds without boundary

In this section we investigate the case where M is $\mathcal{C}^2$, without boundary. Following the definition of our statistical model (Definition 1.15), we consider the class $\mathcal{P}^{2,d,D}_{\tau_{\min},\infty}(f_{\min}, f_{\max})$. From a theoretical viewpoint, convergence rates over this class are well-known, at least in terms of the sample size n.

THEOREM 1.17 :

For $\tau_{\min} > 0$, $0 < f_{\min} \leq f_{\max}$, it holds

(i) [Genovese et al., 2012a, Theorem 3], [Maggioni et al., 2016, Corollary 8], [Divol, 2020, Theorem 3.7]:

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}^{2,d,D}_{\tau_{\min},\infty}(f_{\min},f_{\max})} \mathbb{E}[\mathrm{d_H}(\hat{M}, M)] \leq C_{d,D,\tau_{\min},f_{\min},f_{\max}} \left(\frac{\log(n)}{n}\right)^{\frac{2}{d}},$$

(ii) [Kim and Zhou, 2015, Theorem 1]:

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}^{2,d,D}_{\tau_{\min},\infty}(f_{\min},f_{\max})} \mathbb{E}[\mathrm{d_H}(\hat{M}, M)] \geq c_{d,\tau_{\min},f_{\min},f_{\max}} \left(\frac{\log(n)}{n}\right)^{\frac{2}{d}}.$$

The lower bound is based on the comparison of manifolds with bumped or not areas, in the same spirit as the one used in Section 1.4.3 for Theorem 1.36, Figure 1.15. The upper bound is based on a theoretical estimator, that is an empirical risk minimizer over

the class of $\mathcal{C}^2$ manifolds whose reach is lower bounded by $\tau_{\min}$. Concerning dependency on the sample size n, this result shows that improved convergence rates in $n^{-2/d}$ can be obtained from the empty boundary assumption, compared to the $n^{-1/d}$ rate provided by the (a, d)-standard assumption these distributions satisfy.
In Aamari and Levrard [2018], we propose a constructive estimation scheme based on simplicial complexes. This estimator will also be at use in the case where M has a boundary, as depicted by Theorem 1.25. The construction of this simplicial complex estimator is detailed below.

1.4.1.1 The tangential Delaunay complex estimator

The construction of the tangential Delaunay complex estimator is based on restricted and weighted Delaunay triangulations. We settle here in a deterministic setting, by letting $\mathcal{X}$ be a finite subset of $\mathbb{R}^D$. For $\varepsilon, \delta > 0$, $\mathcal{X}$ is said to be *ε-dense* in M if $\sup_{y\in M} \mathrm{d}(y, \mathcal{X}) \leq \varepsilon$, and *$\delta$-sparse* if $\mathrm{d}(x, \mathcal{X} \setminus \{x\}) \geq \delta$ for all $x \in \mathcal{X}$. A *(δ, ε)-net* (of M) is a δ-sparse and ε-dense point cloud.

Weighted restricted Delaunay Complexes

We now assume that $\mathcal{X} \subset M$. A weight assignment to $\mathcal{X}$ is a function $\omega : \mathcal{X} \longrightarrow [0, \infty)$. The *weighted Voronoi diagram* is defined to be the Voronoi diagram associated to the weighted distance $\mathrm{d}(u, x^\omega)^2 = \|u - x\|^2 - \omega(x)^2$. Every $x \in \mathcal{X}$ is associated to its weighted Voronoi cell $\mathrm{Vor}^\omega(x)$. For $\tau \subset \mathcal{X}$, let

$$\mathrm{Vor}^\omega(\tau) = \bigcap_{x\in\tau} \mathrm{Vor}^\omega(x)$$

be the common face of the weighted Voronoi cells of the points of τ. The *weighted Delaunay triangulation* $\mathrm{Del}^\omega(\mathcal{X})$ is the dual triangulation to the decomposition given by the weighted Voronoi diagram. In other words, for $\tau \subset \mathcal{X}$, the simplex with vertices τ, also denoted by τ, satisfies

$$\tau \in \mathrm{Del}^\omega(\mathcal{X}) \Leftrightarrow \mathrm{Vor}^\omega(\tau) \neq \emptyset.$$

Note that for a constant weight assignment $\omega(x) \equiv \omega_0$, $\mathrm{Del}^\omega(\mathcal{X})$ is the usual Delaunay triangulation of $\mathcal{X}$. Under genericity assumptions on $\mathcal{X}$ and bounds on ω, $\mathrm{Del}^\omega(\mathcal{X})$ is an embedded triangulation with vertex set $\mathcal{X}$ Boissonnat and Ghosh [2014]. The reconstruction method we propose is based on $\mathrm{Del}^\omega(\mathcal{X})$ for some weights ω to be chosen later. As it is a triangulation of the whole convex hull of $\mathcal{X}$ and fails to recover the geometric structure of M, we take restrictions of it in the following manner.
Given a family $R = \{R_x\}_{x\in\mathcal{X}}$ of subsets $R_x \subset \mathbb{R}^D$ indexed by $\mathcal{X}$, the weighted Delaunay complex restricted to R is the sub-complex of $\mathrm{Del}^\omega(\mathcal{X})$ defined by

$$\tau \in \mathrm{Del}^\omega(\mathcal{X}, R) \Leftrightarrow \mathrm{Vor}^\omega(\tau) \cap \left(\bigcup_{x\in\tau} R_x \right) \neq \emptyset.$$

In particular, we define the *Tangential Delaunay Complex* $\mathrm{Del}^\omega(\mathcal{X}, T)$ by taking $R = T = \{T_x M\}_{x\in\mathcal{X}}$, the family of tangent spaces taken at the points of $\mathcal{X} \subset M$ Boissonnat and Ghosh [2014]. $\mathrm{Del}^\omega(\mathcal{X}, T)$ is a pruned version of $\mathrm{Del}^\omega(\mathcal{X})$ where only the simplices with directions close to the tangent spaces are kept. Indeed, $T_x M$ being the best linear approximation of M at x, it is very unlikely for a reconstruction of M to have components in directions normal to $T_x M$ (see Figure 1.6). As pointed out in Boissonnat and Ghosh [2014], computing $\mathrm{Del}^\omega(\mathcal{X}, T)$ only requires to compute Delaunay triangulations in the tangent spaces that have dimension d. This reduces the

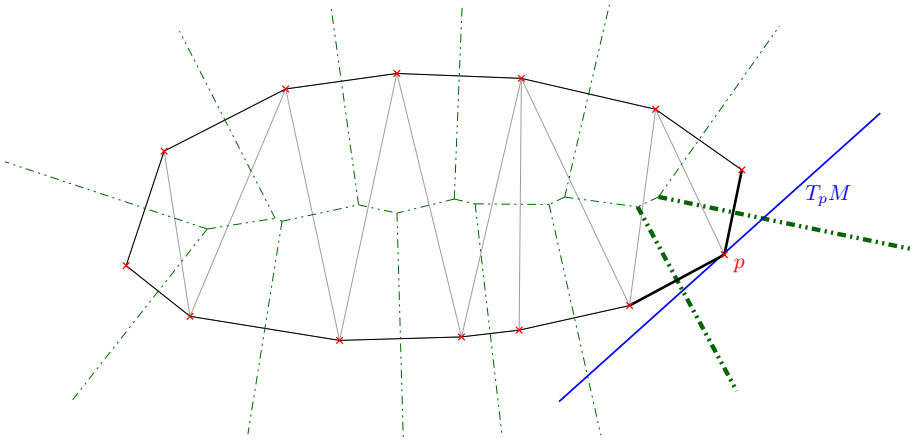


Figure 1.6: Construction of $\mathrm{Del}^{\omega}(\mathcal{X}, T)$ at p for $\omega \equiv 0$: p has three incident edges in the ambient Delaunay triangulation, but only two (bold) have dual Voronoi face intersecting T_pM.

computational complexity dependency on the ambient dimension $D > d$. The weight assignment ω gives degrees of freedom for the reconstruction. The extra degree of freedom ω permits to stabilize the triangulation and to remove the so-called *inconsistencies*, the points remaining fixed. For further details, see Boissonnat et al. [2009], Boissonnat and Ghosh [2014].
Whenever $\mathcal{X} \subset M$ is a $(\varepsilon, 2\varepsilon)$-net of M and $T = \{T_xM\}_{x\in\mathcal{X}}$, reconstruction guarantees of $\mathrm{Del}^{\omega_*}(\mathcal{X}, T)$ are given by [Boissonnat and Ghosh, 2014, Theorem 5.3]. Let us mention that ω_* is a special weight assignment that can be computed from $(\mathcal{X}, T)$ Boissonnat and Ghosh [2014].

Construction of the estimator

To build a weighted tangential Delaunay complex from a random sample point $\mathbb{X}_n$, we have to build a $(\varepsilon, 2\varepsilon)$-net from it, and to give estimates of the tangent spaces. The first point may be adressed using the Farthest-Point Sampling algorithm that originates from graph clustering techniques Gonzalez [1985] and may be described in a few lines: for a point cloud $\mathcal{X}$ and a prescribed resolution ε_0, pick a first point x_1 at random, then select the farthest point x_2 from x_1, the farthest point x_3 from $\{x_1, x_2\}$, etc. Stop when the $k+1$-th farthest point has distance smaller than ε_0 to $\{x_1, \ldots, x_k\}$. It is immediate that, provided $\mathbb{X}_n$ is a ε-covering of M, the Farthest Point Sampling algorithm with scale ε_0 outputs a $(\varepsilon_0, \varepsilon_0 + \varepsilon)$-net of M.
To estimate the tangent spaces at sample points, Local Principal Component Analysis (PCA) is performed. Namely, for $j = 1, \ldots, n$ and a bandwidth $h > 0$, we define the local covariance matrix at $X_j \in \mathbb{X}_n$ by

$$\hat{\Sigma}_j(h) = \frac{1}{n-1} \sum_{i\neq j} \left(X_i - \bar{X}_j\right) \left(X_i - \bar{X}_j\right)^t \mathbb{1}_{\mathrm{B}(X_j,h)}(X_i), \tag{1.2}$$

where $\bar{X}_j = \frac{1}{N_j} \sum_{i\neq j} X_i \mathbb{1}_{\mathrm{B}(X_j,h)}(X_i)$ is the barycenter of sample points contained in the ball $\mathrm{B}(X_j, h)$, and $N_j = |\mathrm{B}(X_j, h) \cap \mathbb{X}_n|$. The estimated tangent space at X_j, $\hat{T}_j$, is then defined as the linear span of the first d eigenvectors of $\hat{\Sigma}_j(h)$.
Our estimator is built as follows: for two parameters $\varepsilon, h > 0$, denote by $\mathbb{Y}_n$ the output of the Farthest Point Sampling algorithm based on $\mathbb{X}_n$, with scale parameter ε, and by $\hat{T}_{\mathbb{Y}_n} = \{\hat{T}_j \mid X_j \in \mathbb{Y}_n\}$. The estimator $\hat{M}_{TDC}$ is defined by

$$\hat{M}_{TDC} = \mathrm{Del}^{\omega_*}(\mathbb{Y}_n, \hat{T}_{\mathbb{Y}_n}). \tag{1.3}$$

THEOREM 1.18 : [AAMARI AND LEVRARD, 2018, THEOREM 6]

Let $P \in \mathcal{P}^{2,d,D}_{\tau_{\min},\infty}(f_{\min}, f_{\max})$, and choose h and ε as follows.

$$h = \left(C_d \frac{f_{\max}^4 \log(n)}{f_{\min}^5 n} \right)^{\frac{1}{d}}, \qquad \varepsilon = \tau_{\min} \left(C_d \frac{f_{\max}^{4+d} \log(n)}{f_{\min}^{5+d} n \tau_{\min}^d} \right)^{\frac{1}{d}}.$$

Then, for n large enough so that $\varepsilon \leq \tau_{\min}/4$, we have, with probability larger than $1 - 2\left(\frac{1}{n}\right)^{\frac{1}{d}}$,

(i) $\mathrm{d_H}(\hat{M}_{TDC}, M) \leq C_d \frac{\varepsilon^2}{\tau_{\min}} \leq C_d \tau_{\min} \left(\frac{f_{\max}^{4+d} \log(n)}{f_{\min}^{5+d} n \tau_{\min}^d} \right)^{\frac{2}{d}}$,

(ii) $\hat{M}_{TDC}$ and M are ambient isotopic.

The bound in terms of Hausdorff distance exposed in Theorem 1.18 has the correct dependency on n (that is $n^{-2/d}$ from Theorem 1.17), and allows to seize the influence of other parameters such as the regularity parameter $\tau_{\min}$. It is worth mentioning that the ambient dimension D plays no role here. Therefore, such a bound is valid uniformly over the class $\mathcal{P}^{2,d,D}_{\tau_{\min},\infty}(f_{\min}, f_{\max})$. The second point of Theorem 1.18 ensures that $\hat{M}_{TDC}$ has the same topology as M. Note however that recovering the topology of M can be performed using simpler estimators such as the Devroye-Wise compact estimator Devroye and Wise [1980], Niyogi et al. [2008]. The interesting fact here is that $\hat{M}_{TDC}$ achieves both minimax optimality over $\mathcal{P}^{2,d,D}_{\tau_{\min},\infty}(f_{\min}, f_{\max})$ and topological consistency. On the technical side, Theorem 1.18 is based on three ingredients: the reconstruction results from [Boissonnat and Ghosh, 2014, Theorem 5.3], standard concentration inequalities for tangent space estimation, and an interpolation lemma that is interesting on its own.

THEOREM 1.19 : [AAMARI AND LEVRARD, 2018, THEOREM 9]

Let $M \in \mathcal{M}^{2,d,D}_{\tau_{\min},\infty}$. Let $\mathcal{X} = \{p_1, \dots, p_q\} \subset \mathbb{R}^D$ be a finite point cloud and $\tilde{T} = \left\{\tilde{T}_1, \dots, \tilde{T}_q\right\}$ be a family of d-dimensional linear subspaces of $\mathbb{R}^D$. For $\theta \leq \pi/64$ and $18\eta < \delta \leq \rho$, assume that

- *$\mathcal{X}$ is δ-sparse: $\min_{i \neq j} \|p_j - p_i\| \geq \delta$,*
- *the p_j's are η-close to M: $\max_{1 \leq j \leq q} \mathrm{d}(p_j, M) \leq \eta$,*
- *$\max_{1 \leq j \leq q} \angle(T_{\pi_M(p_j)} M, \tilde{T}_j) \leq \sin\theta$.*

Then, there exist a universal constant $c_0 \leq 285$ and a compact d-dimensional connected submanifold $M' \subset \mathbb{R}^D$ without boundary such that

1. *$\mathcal{X} \subset M'$,*
2. *$\tau_{M'} \geq \left(1 - c_0\left(\frac{\eta}{\delta} + \theta\right)\frac{\tau_M}{\delta}\right)\tau_M$,*
3. *$T_{p_j} M' = \tilde{T}_j$ for all $1 \leq j \leq q$,*
4. *$\mathrm{d_H}(M, M') \leq \delta\theta + \eta$,*
5. *M and M' are ambient isotopic.*

Theorem 1.19 fits a submanifold M' to noisy points and perturbed tangent spaces with no change of topology and a controlled reach loss. For an appropriate choice of

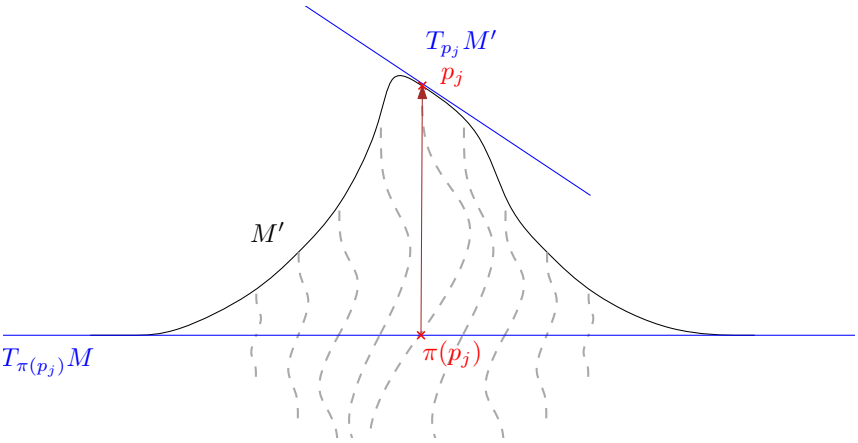


Figure 1.7: An instance of the interpolating submanifold M'. Dashed lines correspond to the image of vertical lines by the ambient diffeomorphism Φ defining $M' = \Phi(M)$.

parameters δ and θ, Theorem 1.19 combined with [Boissonnat and Ghosh, 2014, Theorem 5.3] ensures that $\hat{M}_{TDC}$ is a good interpolation of M', that is both geometrically and topologically close to M. In other words, Theorem 1.19 allows to consider a noisy sample with estimated tangent spaces as an exact sample with exact tangent spaces. The degree of freedom given by η will be of key interest when dealing with boundary estimation, or considering noisy cases. In a nutshell, if ε is chosen as in Theorem 1.18, allowing some orthogonal perturbations of amplitude $\varepsilon^2/\tau_{\min}$ will not deprecate the convergence rate of our Tangential Delaunay Complex estimator $\hat{M}_{TDC}$.

1.4.2 $\mathcal{C}^2$ manifolds with boundary

In this section manifolds are allowed to have boundaries, that is we consider the general model $\mathcal{P}^{2,d,D}_{\tau_{\min},\tau_{\partial,\min}}(f_{\min}, f_{\max})$, for positive $\tau_{\min}$, $\tau_{\partial,\min}$, $f_{\min}$ and $f_{\max}$.
To give an intuition on this problem, let us consider the simple case where M is a d-dimensional convex set of $\mathbb{R}^d$, with $\mathcal{C}^2$ boundary ∂M. Define $\hat{M} = Conv(\mathbb{X}_n)$, and $\widehat{\partial M} = \partial \hat{M}$. Since $\hat{M} \subset M$, we have $\mathrm{d_H}(\hat{M}, M) = \mathrm{d_H}(\partial\hat{M}, \partial M)$, so that the convergence rate for the estimation of M is in fact driven by the convergence rate for the estimation of ∂M. This idea may be found in [Dümbgen and Walther, 1996, Corollary 1] for the convex case, and in [Rodríguez Casal, 2007b, Theorem 3] for the r-convex case. These two results (and the related boundary estimators) rely on the detection and identification of points that are close to the boundary, onto which a boundary estimator is based. We extend this boundary detection approach to the manifold case.

1.4.2.1 Detecting boundary points

In the full-dimensional case $d = D$, under the convexity-type assumptions exposed in Section 1.3, there exists a boundary ∂K, so that the boundary detection problem is trivial. In the submanifold case, a first move is to decide if there exists a boundary, based on the sample $\mathbb{X}_n$. Up to our knowledge, the only result for this question is to be found in Aaron and Cholaquidis [2020], with the additional assumption that the density f is Lipschitz continuous. In this paper, the boundary detection procedure is based on the behavior of local barycenters. In a nutshell, for a point x and bandwidth h, the authors compute the local barycenter $m_x(h) = P_n(du)[u\mathbb{1}_{\mathrm{B}(x,h)}(u)]/P_n(\mathrm{B}(x,h))$, and label x as a boundary point if $\|x - m_x(h)\|$ is large enough (of order $c_0 \times h$). A boundary is then detected if there exists at least one boundary point. Considering the model $\mathcal{P}^{2,d,D}_{\tau_{\min},\tau_{\partial,\min}}(f_{\min}, f_{\max})$, it is likely that this procedure would lead to false discoveries:

indeed, potential jumps of the density lead to displacements of local barycenters that fall above the prescribed threshold (see Figure 1.8).

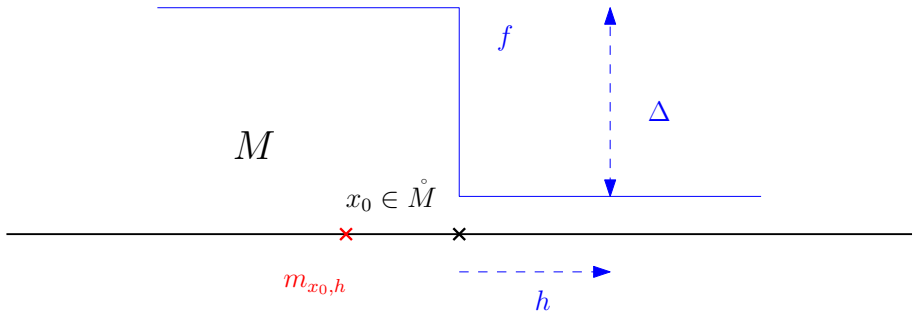


Figure 1.8: One-dimensional case with a jump of the density f at $x_0 \in \mathring{M}$. If $m_{x_0,h}$ denotes the local barycenter of P, easy computation yields $\|m_{x_0,h} - x_0\| \geq \frac{\Delta h}{4f_{\max}}$, that is an order h shift (same order as boundary points).

We rather follow the intuition given in Rodríguez Casal [2007b] in the full-dimensional case, where boundary points are characterized by large Voronoi cells. To be more precise, if $\rho > 0$ denotes a detection radius, the boundary observations may be defined by

$$\mathcal{Y}_\rho = \{X_i \in \mathbb{X}_n \mid \exists O \in \mathbb{R}^D, \|O - X_i\| \geq \rho \text{ and } B(O, \|O - X_i\|) \cap \mathbb{X}_n = \emptyset\}.$$

Interestingly, if $X_i \in \mathcal{Y}_\rho$ for some O in $\mathbb{R}^D$, then $\hat{\eta}_i = \frac{O - X_i}{\|O - X_i\|}$ provides an estimator of the unit outer normal vector at $\pi_{\partial M}(X_i)$ Aaron et al. [2021]. This intuition is depicted by Figure 1.9. For a precise definition of the unit outer normal vector we refer to [Aamari et al., 2021, Proposition 6].
We aim to extend the above intuition to the d-dimensional submanifold case, with $d < D$. First note that even if X_i is far from ∂M, its Voronoi cell can be large in the direction $T_{X_i}M^\perp$. Thus we shall restrict ourselves to the Voronoi cells of the projected sample onto $T_{X_i}M$, or an estimate. Second, if $X_i \in \partial M$ and M is folded over X_i, then the Voronoi cell of X_i in the Voronoi diagram of the projected sample might be small. Thus, not all the sample should be projected but rather $\mathbb{X}_n \cap \bar{\mathrm{B}}(X_i, R_0)$, for some local scale R_0 that will prevent such foldings. Figure 1.10 below illustrates these remarks. These two remarks guide our procedure of boundary points detection.
First, an estimation of tangent spaces is carried out using local PCA again. Namely, we let $\hat{T}_i$ denote the linear span of the first d eigenvectors of $\hat{\Sigma}_j(h)$ (1.2), for $h \lesssim (\log(n)/n)^{1/d}$, and consider the *affine* space $\hat{\mathcal{T}}_i = X_i + \hat{T}_i$. Now, for a local (but macroscopic) scale $R_0 > 0$, a detection radius $\rho > 0$ and a local bandwidth $r > 0$, we compute the Voronoi diagrams of $(\pi_{\hat{\mathcal{T}}_i}(\bar{\mathrm{B}}(X_i, R_0) \cap \mathbb{X}_n))_{i=1,\dots,n}$ and define our boundary points detection procedure as follows.

Definition 1.20 :

Let $i \in [\![1, n]\!]$. We define $J_{R_0,r,\rho}(X_i)$ as the set of r-neighbors of X_i that gives X_i a ρ-large Voronoi cell in the corresponding Voronoi diagram, that is

$$J_{R_0,r,\rho}(X_i) := \{X_j \in \bar{\mathrm{B}}(X_i, r) \cap \mathbb{X}_n \mid \exists O \in \hat{\mathcal{T}}_j \text{ s.t. } \left\|O - \pi_{\hat{\mathcal{T}}_j}(X_i)\right\| \geq \rho \text{ and } \mathrm{B}(O, ||O - \pi_{\hat{\mathcal{T}}_j}(X_i)||) \cap \pi_{\hat{\mathcal{T}}_j}(\bar{\mathrm{B}}(X_j, R_0) \cap \mathbb{X}_n) = \emptyset\}.$$

Then, the set of candidate boundary points $\mathcal{Y}_{R_0,r,\rho}$ is defined as the set of points that have at least one large Voronoi cell (in the collection of Voronoi diagrams), namely

$$\mathcal{Y}_{R_0,r,\rho} = \{X_i \in \mathbb{X}_n \mid J_{R_0,r,\rho}(X_i) \neq \emptyset\}. \tag{1.4}$$

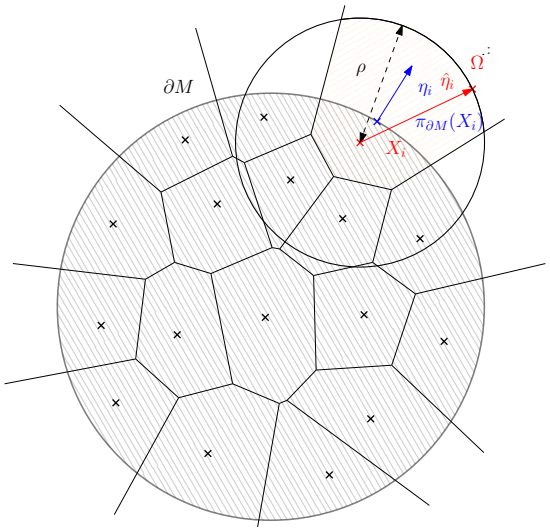


Figure 1.9: In the full-dimensional case, ρ-large Voronoi cells correspond to boundary points, and $\hat{\eta}_i$ provides an estimator of $\eta_{\pi_{\partial M}(X_i)} = \eta_i$.

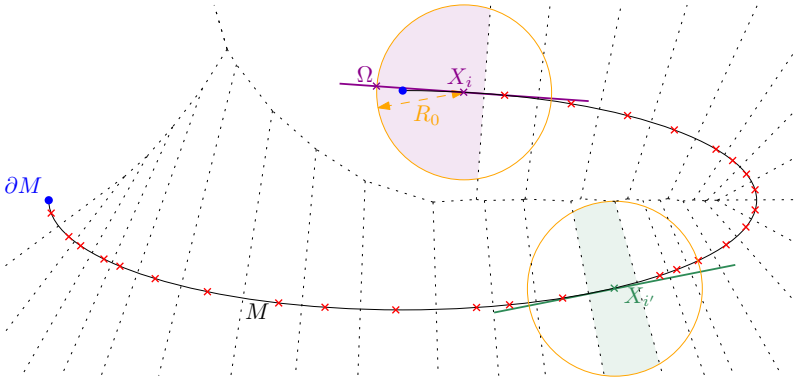


Figure 1.10: An ambient Voronoi diagram built on top of observations $\mathbb{X}_n$ lying on an open plane curve ($d = 1$, $D = 2$). The denser $\mathbb{X}_n$ in M, the narrower the Voronoi cell of the X_i's in the tangent directions $T_{X_i}M$. Observations close to ∂M yield cells that extend in the outward pointing direction. Localization radius $R_0 > 0$ prevents global foldings of M that would mix different ambient neighborhoods of M when projecting onto $T_{X_i}M$.

Detecting boundary points requires to compute n Voronoi diagrams in dimension d. Note that this step does not depend on the ambient dimension D, and can be processed through parallel computation. This boundary points detection procedure also provides a natural way to estimate unit normal outward-pointing vectors, as exposed below.

DEFINITION 1.21 :

Let $X_i \in \mathcal{Y}_{R_0,r,\rho}$, $X_j \in J_{R_0,r,\rho}(X_i)$ and $\Omega_{R_0,r,\rho,j} \in \hat{\mathcal{T}}_j$ be such that $||\Omega_{R_0,r,\rho,j} - \pi_{\hat{\mathcal{T}}_j}(X_i)|| \geq \rho$ and $\mathrm{B}(\Omega_{R_0,r,\rho,j}, \left\|\Omega_{R_0,r,\rho,j} - \pi_{\hat{\mathcal{T}}_j}(X_i)\right\|) \cap \pi_{\hat{\mathcal{T}}_j}(\bar{\mathrm{B}}(X_j, R_0) \cap \mathcal{X}_n) = \emptyset$. The estimator of the unit normal outward-pointing vector in $\hat{T}_j$ is defined by

$$\tilde{\eta}_i^{(j)} = \frac{\Omega_{R_0,r,\rho,j} - \pi_{\hat{\mathcal{T}}_j}(X_i)}{\left\|\Omega_{R_0,r,\rho,j} - \pi_{\hat{\mathcal{T}}_j}(X_i)\right\|}.$$

The final estimator of the unit normal outward-pointing vector at X_i is then defined by

$$\tilde{\eta}_i = \frac{1}{\#J_{R_0,r,\rho}(X_i)} \sum_{X_j \in J_{R_0,r,\rho}(X_i)} \tilde{\eta}_i^{(j)}. \tag{1.5}$$

Theorem 1.22 below provides theoretical guarantees for our detection and normal estimation procedure.

THEOREM 1.22 : [AAMARI ET AL., 2021, THEOREM 19]

Take $R_0 \leq \frac{\tau_{\min} \wedge \tau_{\partial,\min}}{40}$. *Define*

$$r_- := \sqrt{(\tau_{\min} \wedge \tau_{\partial,\min}) R_0} \left(c_d \frac{f_{\max}^5 \log n}{f_{\min}^6 n R_0^d} \right)^{\frac{1}{d+1}}, r_+ := \frac{R_0}{12}, \text{ and } \rho_- := \frac{R_0}{4} =: \frac{\rho_+}{2}$$

Then, for n *large enough, with probability at least* $1 - 4n^{-\frac{2}{d}}$, *we have that for all* $\rho \in [\rho_-, \rho_+]$ *and* $r \in [r_-, r_+]$:

(i) If $\partial M = \emptyset$, *then* $\mathcal{Y}_{R_0,r,\rho} = \emptyset$;

(ii) If $\partial M \neq \emptyset$ *then:*

- *(a) For all* $X_i \in \mathcal{Y}_{R_0,r,\rho}$, $\mathrm{d}(X_i, \partial M) \leq \frac{2r^2}{\tau_{\min} \wedge \tau_{\partial,\min}}$;
- *(b) For all* $x \in \partial M$, $\mathrm{d}(x, \mathcal{Y}_{R_0,r,\rho}) \leq 3r$;
- *(c) For all* $X_i \in \mathcal{Y}_{R_0,r,\rho}$, $\|\eta_{\pi_{\partial M}(X_i)} - \tilde{\eta}_i\| \leq \frac{20r}{\sqrt{R_0(\tau_{\min} \wedge \tau_{\partial,\min})}}$.

The bounds exposed in Theorem 1.22 heavily depend on the scale R_0 and the local bandwidth r, that need to be carefully tuned in practice. Whenever prior information on the reaches $\tau_{\min}$ and $\tau_{\partial,\min}$ is at hand, we may choose R_0 as large as $\frac{\tau_{\min} \wedge \tau_{\partial,\min}}{40}$. Then, an optimal choice $r = r_-$ leads to the bounds:

(ii)a For all $X_i \in \mathcal{Y}_{R_0,r,\rho}$,

$$\mathrm{d}(X_i, \partial M) \leq \tau_{\min} \wedge \tau_{\partial,\min} \left(C_d \frac{f_{\max}^5}{f_{\min}^5} \frac{\log n}{n f_{\min} (\tau_{\min} \wedge \tau_{\partial,\min})^d} \right)^{\frac{2}{d+1}},$$

(ii)b For all $x \in \partial M$,

$$\mathrm{d}\left(x, \mathcal{Y}_{R_0,r,\rho}\right) \leq \tau_{\min} \wedge \tau_{\partial,\min} \left(C_d \frac{f_{\max}^5}{f_{\min}^5} \frac{\log n}{n f_{\min} (\tau_{\min} \wedge \tau_{\partial,\min})^d} \right)^{\frac{1}{d+1}},$$

(ii)c For all $X_i \in \mathcal{Y}_{R_0,r,\rho}$,

$$\|\eta_{\pi_{\partial M}(X_i)} - \tilde{\eta}_i\| \leq \left(C_d \frac{f_{\max}^5}{f_{\min}^5} \frac{\log n}{n f_{\min} (\tau_{\min} \wedge \tau_{\partial,\min})^d} \right)^{\frac{1}{d+1}}.$$

In a nutshell, Item (i) guarantees that no false positive occur if $\partial M = \emptyset$. On the other hand, if $\partial M \neq \emptyset$, for $\varepsilon \asymp (\log n/n)^{1/(d+1)}$ and optimal choices of r_- and R_0, Items (ii)a and (ii)b ensure that $\mathcal{Y}_{R_0,r,\rho}$ is an $O(\varepsilon)$-covering of ∂M that consists of points $O(\varepsilon^2)$-close to ∂M. In the convex case $\tau_{\min} = \infty$, taking the convex hull of $\mathcal{Y}_{R_0,r,\rho}$ — similarly to Dümbgen and Walther [1996] — would result in an $O(\varepsilon^2)$-approximation of M, and the boundary of this convex hull in an $O(\varepsilon^2)$-approximation of ∂M. Finally, Item (ii)c asserts that the estimated normals at boundary observations are $O(\varepsilon)$-precise. To build a boundary estimator, we follow the convex case intuition again. In the convex case, taking the boundary of the convex hull of $\mathcal{Y}_{R_0,r,\rho}$ as in Dümbgen and Walther [1996] results in a ε^2-approximation of ∂M. Still based on $\mathcal{Y}_{R_0,r,\rho}$, we extend this "hull" construction to the non-convex case by leveraging the additional tangential and normal (Theorem 1.22 (ii)c) estimates, to provide estimators of M and ∂M.

1.4.2.2 Boundary estimation

In the general (a, d)-standard case for compact sets included in $\mathbb{R}^d$, Cuevas and Rodríguez-Casal [2004], Aaron and Bodart [2016] shows that the boundary of the Devroye-Wise estimator $\bigcup_{i=1}^n \mathrm{B}(X_i, r_n)$ converges towards ∂K with rate $n^{-1/d}$. Note that, even in this case where boundary estimators are built from compact estimators, the convergence rates of the compact estimator is driven by the boundary convergence rate. When adding regularity conditions, as explained above, a relevant strategy is to build a boundary estimator from detected boundary points. In the $\mathcal{C}^2$ manifold case, a first remark is that ∂M is a $\mathcal{C}^2$ manifold without boundary (Proposition 1.11). Thus, estimators of $\mathcal{C}^2$ boundaryless manifolds such as in Maggioni et al. [2016], Genovese et al. [2012a], Aamari and Levrard [2018] may be a relevant strategy, provided they can adapt to some orthogonal noise. Indeed, on the probability event described by Theorem 1.22, for

$$\varepsilon = C \frac{\tau_{\partial,\min}}{R_0} r,$$

$\mathcal{Y}_{R_0,r,\rho}$ is a ε-covering of ∂M, whose points are $\varepsilon^2/\tau_{\partial,\min}$ close to ∂M. Thus, to apply Theorem 1.19 and provide convergence rates for $\widehat{\partial M}_{TDC}$, it remains to provide boundary tangent space estimators for points in $\mathcal{Y}_{R_0,r,\rho}$.

Definition 1.23 :

> For $X_i \in \mathcal{Y}_{R_0,r,\rho}$, $\hat{T}_{\partial,i}$ is defined as the orthogonal complement of $\pi_{\hat{T}_i}(\tilde{\eta}_i)$ in $\hat{T}_i$. In other words
>
> $$\hat{T}_{\partial,i} := (\pi_{\hat{T}_i}(\tilde{\eta}_i))^{\perp} \cap \hat{T}_i.$$

A straightforward consequence of Theorem [Aamari and Levrard, 2018, Proposition 13] and Theorem 1.22 ensures that $\hat{T}_{\partial,i}$ is a $\varepsilon/\tau_{\partial,\min}$ approximation of $T_{\pi_{\partial M}(X_i)}\partial M$, for any $X_i \in \mathcal{Y}_{R_0,r,\rho}$.

Corollary 1.24 : [Aamari et al., 2021, Corollary 3.10]

Under the assumptions of Theorem 1.22, for a bandwidth h as in Theorem 1.18 for the local PCA, we have, for n large enough, with probability larger than $1-4n^{-\frac{2}{d}}$,

$$\max_{X_i\in\mathcal{Y}_{R_0,r,\rho}} \angle(T_{\pi_{\partial M}(X_i)}\partial M, \hat{T}_{\partial,i}) \leq \frac{20r}{\sqrt{(\tau_{\min}\wedge\tau_{\partial,\min})R_0}}.$$

Interestingly, the estimation rate for $T_{\pi_{\partial M}(X_i)}\partial M$ is driven by $\angle\eta_{\pi_{\partial M}(X_i)},\tilde{\eta}_i$. Equipped with Corollary 1.24, an estimator for ∂M can be build, based on Aamari and Levrard [2018].
For the sake of clarity, we expose results in the case where R_0 is chosen as $\frac{\tau_{\min}\wedge\tau_{\partial,\min}}{40}$ and $r=r_-$. In this framework, we have

$$\varepsilon = C\frac{\tau_{\partial,\min}}{R_0}r_- = \tau_{\partial,\min}\left(c_d\frac{f_{\max}^5\log n}{f_{\min}^6 n(\tau_{\min}\wedge\tau_{\partial,\min})^d}\right)^{\frac{1}{d+1}},$$

and we let $\mathbb{Y}_\partial$ denote a $\varepsilon_{\partial M}$-sparsification of $\mathcal{Y}_{R_0,r,\rho}$ (using Farthest Point Sampling). This results in a $2\varepsilon_{\partial M}$-covering of ∂M, according to Theorem 1.22.
We also denote by $\mathbb{T}_\partial$ the collection of $\hat{T}_{\partial,i}$'s, for $X_i\in\mathbb{Y}_\partial$, and define our estimator of ∂M as the (weighted) Tangential Delaunay Complex based on $(\mathbb{Y}_\partial,\mathbb{T}_\partial)$:

$$\widehat{\partial M} := \mathrm{Del}^{\omega_*}(\mathbb{Y}_\partial,\mathbb{T}_\partial).$$

Since ∂M has no boundary, Theorem 1.19 applies, leading to the following guarantees for boundary estimation based on $\widehat{\partial M}$.

Theorem 1.25 : [Aamari et al., 2021, Theorem 3.11]

Provided that $\partial M\neq\emptyset$ and under the assumptions of Corollary 1.24 with $R_0=\frac{\tau_{\min}\wedge\tau_{\partial,\min}}{40}$ and $r=r_-$, it holds, for n large enough, with probability larger than $1-4n^{-\frac{2}{d}}$,

(i) $\mathrm{d_H}(\partial M,\widehat{\partial M}) \leq \tau_{\partial,\min}\left(C_d\frac{f_{\max}^5\log n}{f_{\min}^6 n(\tau_{\min}\wedge\tau_{\partial,\min})^d}\right)^{\frac{1}{d+1}}$,

(ii) ∂M and $\widehat{\partial M}$ are ambient isotopic.

As a consequence, for n large enough, we have

$$\mathbb{E}\left[\mathrm{d_H}(\partial M,\widehat{\partial M})\right] \leq \tau_{\partial,\min}\left(C_d\frac{f_{\max}^5\log n}{f_{\min}^6 n(\tau_{\min}\wedge\tau_{\partial,\min})^d}\right)^{\frac{2}{d+1}}.$$

As for the case $\partial M=\emptyset$, Theorem 1.25 assesses the topological correctness of our estimator $\widehat{\partial M}$, showing the particular interest of estimators based on simplicial complexes. It also provides a uniform upper bound on $\mathrm{d_H}(\widehat{\partial M},\partial M)$ over the class $\mathcal{P}^{2,d,D}_{\tau_{\min},\tau_{\partial,\min}}(f_{\min},f_{\max})$. This uniform convergence rate is in line with the convergence rate $n^{-2/(d+1)}$ for boundary estimation, under convexity type assumptions in the full dimensional case given in Rodríguez Casal [2007a], Dümbgen and Walther [1996]. Recall that this convex case is a sub-case of our class of distributions $\mathcal{P}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}(f_{\min},f_{\max})$, letting $\tau_{\min}=\infty$. Surprisingly, we can show that this $n^{-2/(d+1)}$ convergence rate is minimax over the class of convex submanifolds.

THEOREM 1.26 : [AAMARI ET AL., 2021, THEOREM 3.12]

Assume that $f_{\min} \leq c_d/\tau_{\partial,\min}^d$ *and* $c'_d/\tau_{\partial,\min}^d \leq f_{\max}$ *for some small enough* $c_d, (c'_d)^{-1} > 0$. *Then for all* $n \geq 1$,

$$\inf_{\hat{B}} \sup_{P \in \mathcal{P}_{\infty,\tau_{\partial,\min}}^{d,D}(f_{\min},f_{\max})} \mathbb{E}_{P^n}\left[\mathrm{d_H}(\partial M, \hat{B})\right] \geq C_d \tau_{\partial,\min} \left\{1 \wedge \left(\frac{1}{f_{\min}\tau_{\partial,\min}^d n}\right)^{2/(d+1)}\right\}.$$

As usual, the proof of Theorem 1.26, relies on Lemma 1.1, the two hypothesis are depicted by Figure 1.11.

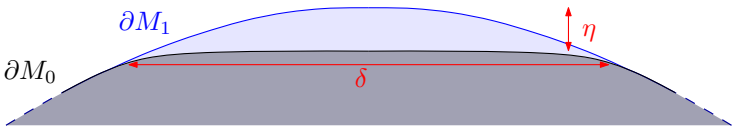


Figure 1.11: Convex supports M_0 and M_1 of Theorem 1.26 for $d = D = 2$. Here, the total variation between the associated uniform distributions is of order $\mathrm{TV}(P_0, P_1) \asymp f_{\min}\mathcal{H}^d(M_0 \triangle M_1) \asymp f_{\min}\delta^{d-1}\eta$ and Hausdorff distance $\mathrm{d_H}(M_0, M_1) = \mathrm{d_H}(\partial M_0, \partial M_1) = \eta$. The reach bound forces the bump to have height $\eta \lesssim \delta^2/\tau_{\partial,\min}$, so that optimal parameter choices yield:

$$\delta \asymp \left(\frac{1}{\tau_{\partial,\min} f_{\min} n}\right)^{1/(d+1)} \quad \text{and} \quad \eta \asymp \frac{\delta^2}{\tau_{\partial,\min}} \asymp \tau_{\partial,\min}\left(\frac{1}{f_{\min}\tau_{\partial,\min}^d n}\right)^{2/(d+1)}.$$

As $\mathrm{TV}(P_0, P_1) \leq 1$, this can only be done when $f_{\min}\delta^{d-1}\eta \lesssim 1$, i.e. $n \gtrsim 1/(f_{\min}\tau_{\partial,\min}^d)$.

Since, for any $\tau_{\min} > 0$, $\mathcal{P}_{\infty,\tau_{\partial,\min}}^{d,D}(f_{\min}, f_{\max}) \subset \mathcal{P}_{\tau_{\min},\tau_{\partial,\min}}^{d,D}(f_{\min}, f_{\max})$, Theorem 1.25 combined with Theorem 1.26 ensures that our boundary estimation procedure is minimax over the class $\mathcal{P}_{\tau_{\min},\tau_{\partial,\min}}^{d,D}(f_{\min}, f_{\max})$, up to $\log n$ factor. Intuitively, these two results show that estimating the boundary under reach conditions on M is not more difficult than estimating the boundary in the convex case, from a statistical viewpoint. Since the convex case is the strongest of the assumptions exposed in Section 1.3, Theorem 1.26 also demonstrates the optimality of the boundary estimators given for the convex case in Dümbgen and Walther [1996], and for the r-convex case in Rodríguez Casal [2007b].

1.4.2.3 Manifold (with boundary) estimation

In the convex case, an estimator of the manifold might be deduced from the boundary estimator, by taking the convex hull. In the manifold case, such an approach can not be readily extended. In general, support estimation methods dedicated to the full-dimensional case, such as the ones exposed in Aaron and Bodart [2016], Rodríguez Casal [2007b], Dümbgen and Walther [1996], cannot be adapted to a low intrinsic dimension framework. For instance, it is easy to see that the r-convex hull described in Rodríguez Casal [2007b] boils down to $\mathbb{X}_n$, for $r \leq \tau_{\min} \wedge \tau_{\partial,\min}$.
We rather adopt a local patches approach, as in Maggioni et al. [2016], Aamari and Levrard [2019]. The points of $\mathbb{X}_n$ will be divided into interior points and boundary points. If X_i is an interior point, then we locally approach M via $\bar{\mathrm{B}}_{\hat{T}_i}(X_i, \varepsilon_{\mathring{M}})$, the affine ball centered at X_i, for some interior radius $\varepsilon_{\mathring{M}}$. Whenever X_i is close the the boundary, it is likely that $\bar{\mathrm{B}}_{\hat{T}_i}(X_i, \varepsilon_{\mathring{M}})$ would spill over ∂M, leading to a poor estimate of M. In this case, replacing $\bar{\mathrm{B}}_{\hat{T}_i}(X_i, \varepsilon_{\mathring{M}})$ by a tangential half-ball with respect to the outward-pointing normal vector $\eta_{\pi_{\partial M}(X_i)}$ seems more appropriate.

We formalize this intuition as follows. Let $\mathcal{Y}_{R_0,r,\rho}$ denote the detected boundary points of Definition 1.20. These points will generate half-balls, with radius $\varepsilon_{\partial M}$, that will roughly approximate the inward slab of radius $\varepsilon_{\partial M}$, $M \cap \bar{\mathrm{B}}(\partial M, \varepsilon_{\partial M})$. To approximate the remaining part of M, we further define the *$\varepsilon_{\partial M}$-inner points* as

$$\mathring{\mathcal{Y}}_{\varepsilon_{\partial M}} := \{X_i \in \mathbb{X}_n \mid \mathrm{d}(X_i, \mathcal{Y}_{R_0,r,\rho}) \geq \varepsilon_{\partial M}/2\}\,. \tag{1.6}$$

Then we define our local patch manifold estimator as follows.

DEFINITION 1.27 :

Given some inner and boundary radii parameters $\varepsilon_{\mathring{M}}$ and $\varepsilon_{\partial M}$, the manifold estimator $\hat{M}$ is defined by

$$\hat{M} := \left(\bigcup_{X_i \in \mathring{\mathcal{Y}}_{\varepsilon_{\partial M}}} \bar{\mathrm{B}}_{\hat{\mathcal{T}}_i}(X_i, \varepsilon_{\mathring{M}}) \right) \cup \left(\bigcup_{X_i \in \mathcal{Y}_{R_0,r,\rho}} \bar{\mathrm{B}}_{\hat{\mathcal{T}}_i}(X_i, \varepsilon_{\partial M}) \cap \{z, \langle z - X_i, \tilde{\eta}_i \rangle \leq 0\} \right),$$

where the $\hat{\mathcal{T}}_i$'s are the affine spaces passing through the X_i's with vector spaces $\hat{T}_i$'s, the estimated tangent spaces via local PCA as in Theorem 1.18. The $\tilde{\eta}_i$'s are the estimators of the outward-pointing normal vectors given by (1.5).

Note that $\hat{M}$ is adaptive in the sense that, if $\partial M = \emptyset$, then, with high probability, we have $\mathcal{Y}_{R_0,r,\rho} = \emptyset$. In this case $\hat{M}$ is the minimax manifold estimator from Aamari et al. [2019] for the class of $\mathcal{C}^2$ manifolds with empty boundaries. Whenever $\partial M \neq \emptyset$, Theorem 1.28 below gives the convergence rate of $\hat{M}$.

THEOREM 1.28 :

Choose (R_0, r, ρ) as in Theorem 1.22, with $R_0 = \frac{\tau_{\min} \wedge \tau_{\partial,\min}}{40}$ and $r = r_-$. Define

$$\varepsilon_{\mathring{M}} = \left(C_d \frac{\log n}{f_{\min} n} \right)^{\frac{1}{d}} \text{ and } \varepsilon_{\partial M} = 18 r_-.$$

Then, for n large enough, it holds

$$\mathbb{E}\left[\mathrm{d_H}(M, \hat{M})\right] \leq C_d \begin{cases} \tau_{\min} \left(\dfrac{f_{\max}^{2+d/2}}{f_{\min}^{2+d/2}} \dfrac{\log n}{f_{\min} \tau_{\min}^d n} \right)^{\frac{2}{d}} & \text{if } \partial M = \emptyset, \\ (\tau_{\min} \wedge \tau_{\partial,\min}) \left(\dfrac{f_{\max}^5}{f_{\min}^5} \dfrac{\log n}{f_{\min} (\tau_{\min} \wedge \tau_{\partial,\min})^d n} \right)^{\frac{2}{d+1}} & \text{if } \partial M \neq \emptyset. \end{cases}$$

Theorem 1.28 is a special case of [Aamari et al., 2021, Theorem 3.14] with the optimal choice of parameters $r = r_-$ and $R_0 = (\tau_{\min} \wedge \tau_{\partial,\min})/40$. In the empty boundary case, $\hat{M}$ achieves the minimax rate $n^{-2/d}$ described in Section 1.4.1. Note however that in the empty boundary case, no guarantees on the topology of $\hat{M}$ may be stated, contrary to the empty boundary dedicated estimator $\hat{M}_{TDC}$ exposed in Theorem 1.18.
Whenever ∂M is not empty, the convergence rate of our estimator is in fact driven by the convergence rate of the boundary estimation problem of Theorem 1.25, as for the convex and r-convex full-dimensional case. Note that the convergence rates for these two cases, given in [Dümbgen and Walther, 1996, Corollary 1] for the convex case and in

[Rodríguez Casal, 2007a, Theorem 3] for the r-convex case, are recovered by Theorem 1.28, at least in terms of sample size dependency. As for the boundary estimation problem, we can show that this $n^{-2/(d+1)}$ rate is in fact the minimax rate over the class of convex $\mathcal{C}^2$ submanifolds.

THEOREM 1.29 :

(i) (Boundariless Case): Assume that $f_{\min} \leq c_d/\tau_{\min}^d$ and $c'_d/\tau_{\min}^d \leq f_{\max}$, for some small enough $c_d, (c'_d)^{-1} > 0$. Then for all $n \geq 1$,

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}^{d,D}_{\tau_{\min},\infty}(f_{\min},f_{\max})} \mathbb{E}_{P^n}\left[\mathrm{d_H}(M,\hat{M})\right] \geq C_d \tau_{\min} \left\{1 \wedge \left(\frac{1}{f_{\min}\tau_{\min}^d n}\right)^{2/d}\right\}.$$

(ii) (Convex Case): Assume that $f_{\min} \leq c_d/\tau_{\partial,\min}^d$ and $c'_d/\tau_{\partial,\min}^d \leq f_{\max}$, for some small enough $c_d, (c'_d)^{-1} > 0$. Then for all $n \geq 1$,

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}^{d,D}_{\infty,\tau_{\partial,\min}}(f_{\min},f_{\max})} \mathbb{E}_{P^n}\left[\mathrm{d_H}(M,\hat{M})\right] \geq C_d \tau_{\partial,\min} \left\{1 \wedge \left(\frac{1}{f_{\min}\tau_{\partial,\min}^d n}\right)^{2/(d+1)}\right\}.$$

The hypothesis used in the proof of Theorem 1.29, Item (ii) are the same as in the proof of Theorem 1.26. The first point is a slight refinement of [Aamari et al., 2019, Theorem 7] or [Genovese et al., 2012b, Theorem 2], that exhibits the dependency on $\tau_{\min}$ and $f_{\min}$ of the minimax rates over the class of $\mathcal{C}^2$ manifolds with no boundary. Interestingly, this shows that the upper bounds given in Theorem 1.28 or Theorem 1.18 for the empty boundary case are sharp with respect to $\tau_{\min}$. The second point of Theorem 1.29 provides the minimax rate for manifold estimation over the class of convex manifolds whose boundary has bounded reach. Concerning the dependency on the sample size, this shows that our estimator has the correct $n^{-2/(d+1)}$ convergence rate for this convex case, as well as the two aforementioned procedures of Dümbgen and Walther [1996], Rodríguez Casal [2007a]. Note however that [Korostelëv and Tsybakov, 1994, Theorem 4] provides a lower bound in $(\log(n)/n)^{2/(d+1)}$ in the convex and $d = 2$ case. This suggests that the sample size dependency may be improved in Theorem 1.29, borrowing techniques from Kim and Zhou [2015] for instance.
As for the boundary estimation problem, this result intuitively carries the message that estimating a manifold with boundary under reach conditions is not more difficult than estimating a d-dimensional convex $\mathcal{C}^2$-domain. In other words, for $\partial M \neq \emptyset$ and a fixed boundary's convexity radius $\tau_{\partial,\min}$, no additional gain can be expected from requiring a large convexity radius for the manifold (driven by $\tau_{\min}$). At last, Theorem 1.28 shows that the given dependency on the reach boundary $\tau_{\partial,\min}$ is sharp, at least in the case where $\tau_{\partial,\min} \leq \tau_{\min}$. Whether the tradeoff between $\tau_{\min}$ and $\tau_{\partial,\min}$ exhibited in Theorem 1.28 is sharp in general remains an open question.

1.4.3 Smooth(er) manifolds without boundary

The previous sections show that a controlled regularity of order 2 for M (and ∂M) provides improved convergence rates for the support estimation problem, of order $n^{-2/d}$ in the empty boundary case and $n^{-2/(d+1)}$ if the boundary is not empty. A natural question is to determine whether requiring more smoothness can further improve the convergence rates. The intuition here is simple: consider a 1-dimensional manifold M in $\mathbb{R}^2$, going through 0 and whose tangent space at 0 is not $\mathbb{R}e_2$ (the vertical axis), see

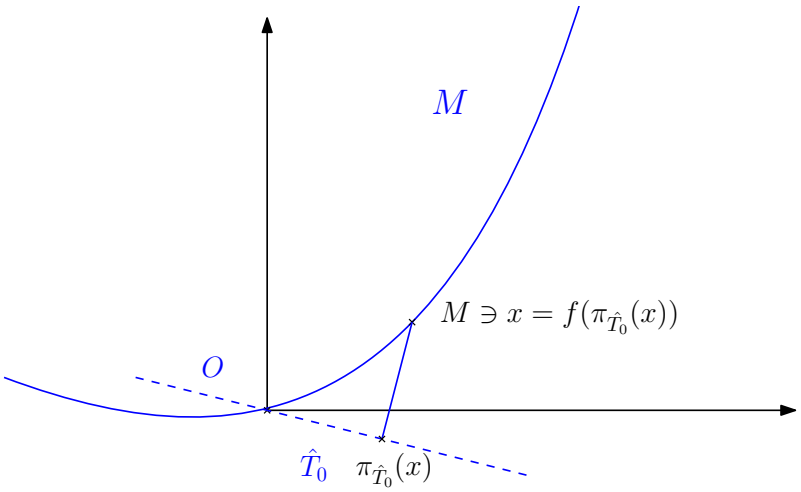


Figure 1.12: One-dimensional manifolds can be locally parametrized by their projections on estimated tangent spaces.

Figure 1.12. Then, around 0, M is the graph of a function f.

The estimators exposed in Section 1.4.2 and 1.4.1 consist in locally approximating f with a linear term around 0, in terms of the local coordinates defined by $\hat{T}_0$. Provided the estimated tangent space is not $\mathbb{R}e_2$, for a scale h, this provides a h^2-approximation of M near 0. Now if f is assumed to be $\mathcal{C}^k$ around 0, then fitting a polynomial order $k-1$ in terms of the local coordinates could lead to a h^k approximation of M, locally around 0. This approach can be extended to arbitrary intrinsic dimension d, for instance using osculating d-jets as in Cazals and Pouget [2005], provided that around every point, the manifold can be locally expressed as the graph of a $\mathcal{C}^k$ map for some coordinate system.

1.4.3.1 An appropriate $\mathcal{C}^k$ model

We investigate here the case where M has no boundary. According to Definition 1.9, if M is a $\mathcal{C}^k$ manifold, then locally around each $p \in M$, M can locally be expressed as the image of a $\mathcal{C}^k$ function with arguments in T_pM (for a scale h small enough so that $\pi_{T_pM} \circ \Psi_p : \left(U_0 \cap \left(\mathbb{R}^d \times \{0\}^{D-d}\right)\right) \to \pi_{T_p}\left[M \cap V_p \cap B(p,h)\right]$ is a diffeomorphism). We may then parametrize our regularity class by:

(i) Requiring that M can be locally parametrized by $\mathcal{C}^k$ diffeomorphisms, for a local scale that is uniform over M (such a scale exists, by compactness arguments).

(ii) Requiring bounds on the p-th derivatives of such parametrizations, $p = 1, \ldots, k$, uniformly on M. This prevents jumps into the $\mathcal{C}^{k-1}$ class when considering uniform bounds over the $\mathcal{C}^k$ model.

This lead us to introduce the following $\mathcal{C}^k$ regularity class.

Definition 1.30 :

For $k \geq 3$, $\tau_{\min} > 0$, and $\mathbf{L} = (L_\perp, L_3, \ldots, L_k)$, we let $\mathcal{M}^{k,d,D}_{\tau_{\min},\infty,\mathbf{L}}$ denote the set of d-dimensional compact connected submanifolds M of $\mathbb{R}^D$ with $\tau_M \geq \tau_{\min}$ and such that, for all $p \in M$, there exists a local one-to-one parametrization Ψ_p of the form:

$$\begin{aligned} \Psi_p \colon \mathrm{B}_{T_pM}(0,r) &\longrightarrow M \\ v &\longmapsto p + v + \mathbf{N}_p(v) \end{aligned}$$

for some $r \geq \frac{1}{4L_\perp}$, with $\mathbf{N}_p \in \mathcal{C}^k\left(\mathrm{B}_{T_pM}(0,r), \mathbb{R}^D\right)$ such that

$$\mathbf{N}_p(0) = 0, \quad d_0\mathbf{N}_p = 0, \quad \left\|d^2_v \mathbf{N}_p\right\|_{op} \leq L_\perp,$$

for all $\|v\| \leq \frac{1}{4L_\perp}$. Furthermore, we require that

$$\left\|d^i_v \mathbf{N}_p\right\|_{op} \leq L_i \text{ for all } 3 \leq i \leq k.$$

For the case $k = 2$, [Aamari and Levrard, 2019, Lemma 1] shows that the exponential parametrization roughly satisfies the requirements of Definition 1.30, with $L_\perp = 1/\tau_{\min}$. From this point of view, Definition 1.30 extends the relevant properties of the exponential map in the $\mathcal{C}^2$ case to the $\mathcal{C}^k$ case, $k \geq 3$. Note that for $k \geq 3$ the exponential map can happen to be only $\mathcal{C}^{k-2}$ for a $\mathcal{C}^k$-submanifold Hartman [1951]. Hence, it may not be a good choice of Ψ_p.

It is important to note that such a family of Ψ_p's exists for any compact $\mathcal{C}^k$-submanifold, if one allows $\tau^{-1}_{\min}$, $L_\perp$, L_3,...,L_k to be large enough. The radius $1/(4L_\perp)$ has been chosen for convenience, other smaller scales would do and we could even parametrize this constant, but without substantial benefits in the results.

The Ψ_p's can be seen as unit parametrizations of M. The conditions on $\mathbf{N}_p(0)$, $d_0\mathbf{N}_p$, and $d^2_v\mathbf{N}_p$ ensure that Ψ_p^{-1} is close to the projection π_{T_pM} (like the exponential map in the $\mathcal{C}^2$-case, see [Aamari and Levrard, 2019, Lemma 1]). The bounds on $d^i_v\mathbf{N}_p$ $(3 \leq i \leq k)$ allow to control the coefficients of the polynomial expansion we seek. Indeed, whenever $M \in \mathcal{C}^k_{\tau_{\min},\mathbf{L}}$, we can show that for every p in M, and y in $\mathrm{B}\left(p, \frac{\tau_{\min} \wedge L_\perp^{-1}}{4}\right) \cap M$,

$$y - p = \pi^*(y-p) + \sum_{i=2}^{k-1} T^*_i(\pi^*(y-p)^{\otimes i}) + R_k(y-p), \tag{1.7}$$

where π^* denotes the orthogonal projection onto T_pM, the T^*_i are i-linear maps from T_pM to $\mathbb{R}^D$ with $\|T^*_i\|_{op} \leq L'_i$ and R_k satisfies $\|R_k(y-p)\| \leq C\|y-p\|^k$, where the constants C and the L'_i's depend on the parameters $\tau_{\min}$, d, k, $L_\perp, \ldots, L_k$. The technical details that relates the L'_i's to the parameter of the model $\mathcal{M}^{k,d,D}_{\tau_{\min},\infty,\mathbf{L}}$ can be found in [Aamari and Levrard, 2019, Lemma A.2, Appendix].

Considering the $\mathcal{M}^{k,d,D}_{\tau_{\min},\infty,\mathbf{L}}$, a natural question is to determine whether it is over-parametrized, in the sense that controlling $\mathbf{L} = (L_\perp, L_3, \ldots, L_k)$ could be enough to lower bound $\tau_{\min}$. A first step is to recall that the reach allows to control the curvature of the manifold, but also the size of bottlenecks. Following this intuition and Aamari et al. [2019], we may define the *local* reach and *global* reach as follows.

Definition 1.31 :

Let M be a compact $\mathcal{C}^2$ submanifold of $\mathbb{R}^D$ without boundary. The local reach R_ℓ is defined by

$$R_\ell = \min_{x \in M} \|\mathrm{II}_x\|_{op}^{-1},$$

where we recall that II_x is the second fundamental form of M at x.

The global reach R_g is defined by

$$R_g = \min\{\|q_1 - q_2\| \mid q_1, q_2 \in M \quad \text{and} \quad (q_1 + q_2)/2 \in \mathrm{Med}(M)\},$$

where $\mathrm{Med}(M)$ denotes the medial axis of M defined in (1.1).

Note that the global reach R_g is closely related to the weak feature size, that characterizes the distance from M to the set of critical points of the distance function d_M (see, e.g., Attali et al. [2013] for a formal definition). Using Proposition 1.12, it is immediate that $\tau_M \leq R_g \wedge R_\ell$. The following result from Aamari et al. [2019] guarantees that the local and global reach fully characterize τ_M.

Theorem 1.32 : [Aamari et al., 2019, Theorem 3.4]

Let $M \subset \mathbb{R}^D$ be a compact manifold without boundary. Then

$$\tau_M = R_\ell \wedge R_g.$$

Equipped with Theorem 1.32, we can now conceive $\mathcal{C}^k$ manifolds with fixed smoothness parameters $\mathbf{L}$ whose reach tends to 0, by letting R_g decrease. To do so, start from the curves depicted in Figure 1.13.

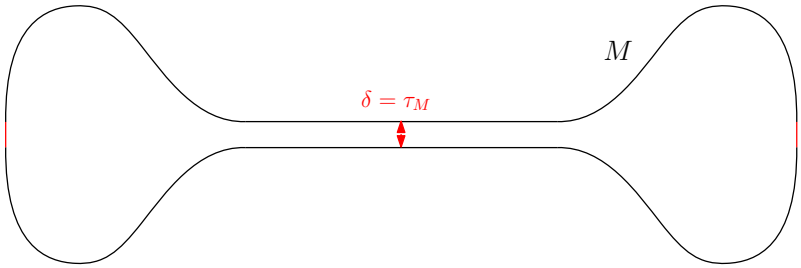


Figure 1.13: One dimensional manifolds with fixed $\mathbf{L}$, but $\tau_M = \delta \to 0$.

To obtain d-dimensional manifolds, we may consider Cartesian products of these curves, that are embedded in $\mathbb{R}^{2d}$. This provides us with manifolds $M \in \mathcal{M}^{k,d,2d}_{0,\infty,\mathbf{L}}$, for a fixed $\mathbf{L}$, whose reach is arbitrarily small, addressing the over-parametrization issue.

We may now define our statistical model $\mathcal{P}^{k,d,D}_{\tau_{\min},\infty,\mathbf{L}}(f_{\min}, f_{\max})$, that slightly differs from Definition 1.15.

Definition 1.33 :

For $k \geq 3$, $\tau_{\min} > 0$, $\mathbf{L} = (L_\perp, L_3, \dots, L_k)$ and $0 < f_{\min} \leq f_{\max}$, we define $\mathcal{P}^{k,d,D}_{\tau_{\min},\infty,\mathbf{L}}(f_{\min}, f_{\max})$ as the set of Borel probability distributions P on $\mathbb{R}^D$ such that:

- $M = \mathrm{Supp}(P) \in \mathcal{M}^{k,d,D}_{\tau_{\min},\infty,\mathbf{L}}$,
- P has a density f with respect to the volume measure $\mathrm{vol}_M = \mathbb{1}_M \mathcal{H}^d$ on M, such that $f_{\min} \leq f(x) \leq f_{\max}$ for all $x \in M$.

It is immediate that $\mathcal{P}^{k,d,D}_{\tau_{\min},\infty,\mathbf{L}}(f_{\min}, f_{\max}) \subset \mathcal{P}^{k,d,D}_{\tau_{\min},\infty}(f_{\min}, f_{\max})$, and conversely, for any P in $\mathcal{P}^{k,d,D}_{\tau_{\min},\infty}(f_{\min}, f_{\max})$, there exists $\mathbf{L}$ such that $P \in \mathcal{P}^{k,d,D}_{\tau_{\min},\infty,\mathbf{L}}(f_{\min}, f_{\max})$.
Proposition 1.16 ensures that if $P \in \mathcal{P}^{k,d,D}_{\tau_{\min},\infty,\mathbf{L}}(f_{\min}, f_{\max})$, then P is $(c_d f_{\min}, d)$-standard at scale $\tau_{\min}/16$. Since the $\mathcal{M}^{k,d,D}_{\tau_{\min},\infty,\mathbf{L}}$ model allows to separately investigate local and global regularity, a more refined result might be stated.

Proposition 1.34 :

Let $P \in \mathcal{P}^{k,d,D}_{\tau_{\min},\infty,\mathbf{L}}(f_{\min}, f_{\max})$. Then P is $(c_d f_{\min}, d)$-standard at scale $1/(4L_\perp)$.

Proof of Proposition 1.34. Let $p \in M$. For $u \in T_pM$ such that $\|u\| \leq 1/(4L_\perp)$, and $t \in [0,1]$, denote by $g(t) = d_{tu}N_p$. According to Definition 1.33, we may write

$$\begin{aligned} d_u N_p &= d_0 N_p + \int_0^1 g'(t) dt \\ &= d_0 N_p + \int_0^1 d^2 N_{tu}(u) dt, \end{aligned}$$

so that $\|d_u N_p\|_{op} \leq L_\perp \|u\|$. Proceeding as above leads to $\|N_p(u)\| \leq \frac{L_\perp \|u\|^2}{2}$, that entails

$$\frac{7}{8}\|u\| \leq \|\psi_p(u) - p\| \leq \frac{9}{8}\|u\|,$$

as well as

$$J_u \psi_p \geq (1 - \|d_u N_p\|_{op})^d \geq \left(\frac{3}{4}\right)^d,$$

where $J_u\psi_p$ denotes the Jacobian of ψ_p at u. Now let $h \leq 1/(4L_\perp)$. We may write

$$\begin{aligned} \int_M \mathbb{1}_{\mathrm{B}(p,h)}(x) f(x) \,\mathrm{vol}_M(dx) &\geq \int_M \mathbb{1}_{\psi_p(\mathrm{B}_{T_pM}(0,8h/9))}(x) f(x) \,\mathrm{vol}_M(dx) \\ &\geq f_{\min} \int_{\mathrm{B}_{T_pM}(0,8h/9)} J\psi_p(u) du \\ &\geq c_d f_{\min} h^d, \end{aligned}$$

using the co-area formula ([Federer, 1959, Theorem 3.1]). □

An interesting consequence of Proposition 1.34 is that $\mathcal{P}^{k,d,D}_{0,\infty,\mathbf{L}}(f_{\min}, f_{\max})$ is a subset of $(c_d f_{\min}, d)$-standard measures at scale $1/(4L_\perp)$, that guarantees a uniform convergence rate of $(1/n)^{1/d}$ for Hausdorff estimation over this class, according to Theorem 1.4. In other words, letting $\tau_{\min}$ be arbitrarily small with fixed smoothness parameters $\mathbf{L}$ still entails consistency of some manifold estimators. This is not the case for tangent space and curvature estimation, as alleged by [Aamari and Levrard, 2019, Theorem 1].
Now, for a distribution $P \in \mathcal{P}^{k,d,D}_{\tau_{\min},\infty,\mathbf{L}}(f_{\min}, f_{\max})$, we may formalize our local polynomial intuition.

1.4.3.2 Local polynomials

For this, we need one more piece of notation. For $1 \leq j \leq n$, $P_{n-1}^{(j)}$ denotes integration with respect to $1/(n-1)\sum_{i\neq j}\delta_{(X_i - X_j)}$, and $z^{\otimes i}$ denotes the $D \times i$-dimensional vector $(z, \dots, z)$. For a bandwidth $h > 0$ to be chosen later, we define the local polynomial estimator $(\hat{\pi}_j, \hat{T}_{2,j}, \dots, \hat{T}_{k-1,j})$ at X_j to be any element of

$$\underset{\pi, \sup_{2\leq i\leq k}\|T_i\|_{op}\leq 1/h}{\arg\min} P_{n-1}^{(j)}\left[\left\|x - \pi(x) - \sum_{i=2}^{k-1} T_i(\pi(x)^{\otimes i})\right\|^2 \mathbb{1}_{\mathrm{B}(0,h)}(x)\right], \tag{1.8}$$

where π ranges among all the orthogonal projectors on d-dimensional subspaces, and $T_i : \left(\mathbb{R}^D\right)^i \to \mathbb{R}^D$ among the symmetric tensors of order i such that $\|T_i\|_{op} \leq 1/h$. Note that compactness of the domain of minimization ensures that such a minimizer exists almost surely. The bound $1/h$ on $\|T_i\|_{op}$ follows from technical considerations, that are detailed in [Aamari and Levrard, 2019, Theorem 2]. Such a constraint on the higher order tensors might have been stated under the form of a $\|.\|_{op}$-penalized least squares minimization (as in ridge regression) leading to the same results.
In the case where $k = 2$, the sum over the tensors T_i is empty, and the integrated term reduces to $\|x - \pi(x)\|^2 \mathbb{1}_{\mathrm{B}(0,h)}(x)$. We thus recover the local PCA tangent spaces estimates defined below (1.2), that eventually lead to a local linear patch estimator of M as in Section 1.4.2.
From a computational viewpoint, it is worth noting that the set of d-dimensional orthogonal projectors is not convex, which leads to a more involved optimization problem than usual least squares. In practice, this problem may be solved using tools from optimization on Grassman manifolds Usevich and Markovsky [2014], or adopting a two-stage procedure such as in Cazals and Pouget [2005]: from local PCA, a first d-dimensional space is estimated at each sample point, along with an orthonormal basis of it. Then, the optimization problem (1.8) is expressed as a minimization problem in terms of the coefficients of $(\pi_j, T_{2,j}, \dots, T_{k,j})$ in this basis under orthogonality constraints.
Before turning to the definition of our $\mathcal{C}^k$ manifold estimator, let us mention that some coefficients of $(\hat{\pi}_j, \hat{T}_{2,j}, \dots, \hat{T}_{k-1,j})$ are of particular interest: indeed, choosing $\hat{T}_j$ as $\operatorname{im}\hat{\pi}_j$ provides a tangent space estimator that has convergence rate $n^{-(k-1)/d}$, according to [Aamari and Levrard, 2019, Theorem 2]. Further, when $k \geq 3$, letting $\hat{\mathrm{II}}_j = \hat{T}_{2,j} \circ \hat{\pi}_j$ provides an estimator of II_{X_j} that has convergence rate $n^{-(k-2)/d}$ ([Aamari and Levrard, 2019, Theorem 4]). In a nutshell, several geometric quantities on M may be inferred from the coefficients $(\hat{\pi}_j, \hat{T}_{2,j}, \dots, \hat{T}_{k-1,j})$.
Now, for a fixed $j \in [\![1, n]\!]$, we try to estimate the local parametrization exposed in (1.7) by defining

$$\hat{\Psi}_j(v) = X_j + v + \sum_{i=2}^{k-1} \hat{T}_{i,j}\left(v^{\otimes i}\right),$$

for $v \in \operatorname{im}\hat{\pi}_j$, where $(\hat{\pi}_j, \hat{T}_{2,j}, \dots, \hat{T}_{k-1,j})$ are given by (1.8). For a bandwidth h, our manifold estimator $\hat{M}$ can then be defined as

$$\hat{M} = \bigcup_{j=1}^{n} \hat{\Psi}_j\left(\mathrm{B}_{\hat{T}_j}(0, 7h/8)\right), \tag{1.9}$$

where $\hat{T}_j = \operatorname{im}\hat{\pi}_j$. Contrary to the estimator $\hat{M}_{TDC}$ given in Theorem 1.18 for the $\mathcal{C}^2$ case, the set $\hat{M}$ has no reason to be globally smooth, since it consists of a mere union of polynomial patches (see Figure 1.14). However, $\hat{M}$ is provably close to M for the Hausdorff distance.

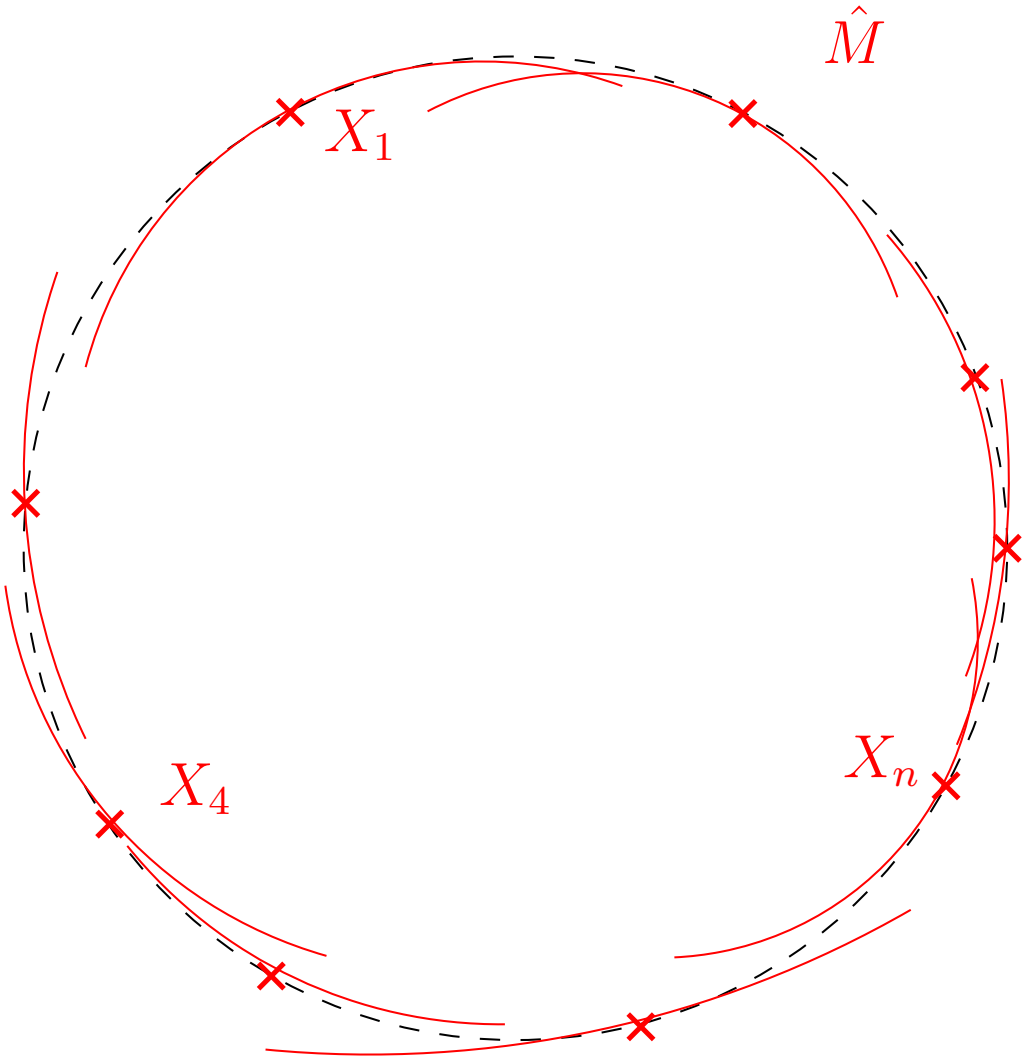


Figure 1.14: A one-dimensional quadratic patches estimator, that may not even be connected.

THEOREM 1.35 :

Let $P \in P \in \mathcal{P}^{k,d,D}_{\tau_{\min},\infty,\mathbf{L}}(f_{\min}, f_{\max})$, *and* $h = \left(C_{d,k}\frac{f_{\max}^2 \log(n)}{f_{\min}^3 (n-1)}\right)^{\frac{1}{d}}$. *Then, for* n *large enough so that* $h \leq (\tau_{\min} \wedge L_\perp^{-1})/8$, *we have, with probability at least* $1 - 2\left(\frac{1}{n}\right)^{\frac{k}{d}}$,

$$\mathrm{d_H}(M, \hat{M}) \leq C_{d,k,\tau_{\min},\mathbf{L}} \sqrt{\frac{f_{\max}}{f_{\min}}} h^k.$$

In particular, for n *large enough,*

$$\sup_{P \in \mathcal{P}^{k,d,D}_{\tau_{\min},\infty,\mathbf{L}}(f_{\min},f_{\max})} \mathbb{E}_{P^{\otimes n}} \mathrm{d_H}(M, \hat{M}) \leq C_{d,k,\tau_{\min},\mathbf{L}} \left(\frac{f_{\max}^{2+\frac{d}{2k}} \log n}{f_{\min}^{3+\frac{d}{2k}} (n-1)}\right)^{\frac{k}{d}}.$$

For greater orders of smoothness ($k \geq 3$), the local polynomial estimator $\hat{M}$ outperforms reconstruction procedures based on a somewhat piecewise linear interpolation Aamari and Levrard [2018], Genovese et al. [2012a], Maggioni et al. [2016], and achieves the

faster rate $(\log n/n)^{k/d}$ for the Hausdorff loss. This is done at the price of a probably worse dependency on the dimension d than in Aamari and Levrard [2018], Genovese et al. [2012a]. This faster rate of convergence may be proved (almost) minimax optimal over the class $\mathcal{P}^{k,d,D}_{\tau_{\min},\infty,\mathbf{L}}(f_{\min}, f_{\max})$.

THEOREM 1.36 :

If $\tau_{\min} L_\perp, \dots, \tau_{\min}^{k-1} L_k, (\tau_{\min}^d f_{\min})^{-1}$ and $\tau_{\min}^d f_{\max}$ are large enough (depending only on d and k), then for n large enough,

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}^{k,d,D}_{\tau_{\min},\infty,\mathbf{L}}(f_{\min},f_{\max})} \mathbb{E}_{P^{\otimes n}} \mathrm{d_H}(M, \hat{M}) \geq c_{d,k,\tau_{\min}} \left(\frac{1}{n}\right)^{\frac{k}{d}},$$

where the infimum is taken over all the estimators $\hat{M} = \hat{M}(X_1, \dots, X_n)$.

The proof of Theorem 1.36 is based on Lemma 1.1 with suitable hypothesis depicted in Figure 1.15.

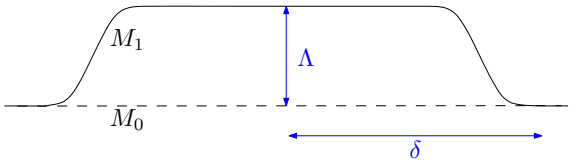


Figure 1.15: Manifolds M_0 and M_1 of Theorem 1.36. The width $\delta \asymp (\tau_{\min} f_{\min} n)^{-\frac{1}{d}}$ of the bump is chosen to have $\mathrm{TV}(P_0, P_1) \leq \frac{1}{n}$. The distance $\Lambda = \mathrm{d_H}(M_0, M_1)$ is of order δ^k to ensure that $M_1 \in \mathcal{C}^k_{\tau_{\min},\infty}(f_{\min}, f_{\max})$.

Let us note that it is likely for the extra $\log n$ term appearing in Theorem 1.35 to actually be present in the minimax rate. Roughly, it is due to the fact that the Hausdorff distance $\mathrm{d_H}$ is similar to a L^∞ loss. The $\log n$ term might be obtained in Theorem 1.36 with the same combinatorial analysis as in Kim and Zhou [2015] for $k = 2$. In some sense, Theorem 1.36 generalizes Theorem 1.17, *ii*) to higher orders of regularity ($k \geq 3$). Unfortunately, the constants $C_{d,k,\tau_{\min},\mathbf{L}}$ and $c_{d,k,\tau_{\min},\mathbf{L}}$ in Theorems 1.35 and 1.36 do not allow to seize the influence of the regularity parameters $\tau_{\min}$ and $\mathbf{L}$. However, some intuition on the importance of these parameters may be given. First, if $\mathbf{L}$ is allowed to be arbitrarily large and $\tau_{\min}$ arbitrarily small, then adding a small sphere of radius $\tau_{\min} \sim n^{-1/d}$ far from a $\mathcal{C}^k$ manifold render the manifold estimation problem undecidable, leading to a constant lower bound over $\mathcal{P}^{k,d,D}_{0,\infty}(f_{\min}, f_{\max})$. Note that in this case both $(L_\perp)^{-1}$ and $\tau_{\min}$ tend to 0. It is clear that a high curvature, encoded by $L_\perp$, entails a small reach, thus letting $L_\perp$ tend to ∞ will imply $\tau_{\min} \to 0$.
As alleged by Proposition 1.12 and Figure 1.13, the converse is not true, and a natural question is to determine which convergence rate may be expected on $\mathcal{P}^{k,d,D}_{0,\infty,\mathbf{L}}(f_{\min}, f_{\max})$, for a fixed $\mathbf{L}$. Combining Proposition 1.34 with Theorem 1.4 ensures that the convergence rate over $\mathcal{P}^{k,d,D}_{0,\infty,\mathbf{L}}(f_{\min}, f_{\max})$ should be smaller than $(1/n)^{\frac{1}{d}}$. Nevertheless, up to our knowledge there exists no lower bound to match this rate. Its optimality remains an open question.
An other open question is to determine the minimax convergence rate over the class $\mathcal{P}^{k,d,D}_{\tau_{\min},\tau_{\partial,\min}}(f_{\min}, f_{\max})$, that is over the class of $\mathcal{C}^k$ manifolds with boundary. We conjecture that this rate is $n^{-k/(k-1+d)}$, under suitable smoothness assumptions on both M and ∂M. Indeed, for a positive h, take a h-cover of ∂M, and consider the "slabs" of

heights h^k and basis h, that is, for an element x of the h-covering,

$$S_x(h) = \{y \in M \mid d(y, \partial M) \leq h^k \quad \text{and} \quad \mathrm{d}_{\partial M}(\pi_{\partial M}(y), x) \leq h\}.$$

Since the volume of such a slab is roughly $h^{d-1} \times h^k$, the probability that $\mathbb{X}_n$ leave such a slab empty can be made polynomial in n^{-1} by choosing $h \lesssim (\log(n)/n)^{1/(k+d-1)}$. Thus, were we able to identify those (very) close to boundary points, then estimating the boundary with rate h^k would be possible using a local polynomial patch estimator of ∂M and [Aamari and Levrard, 2019, Theorem 6] (that allows for small orthogonal perturbations). Therefore, the key point would be to detect those h^k-close points to ∂M, that remains out of our reach for the moment.
Before turning to noisy situations, let us recap the exposed convergence rates for support estimation in the noiseless case, depending on support smoothness.

	(a,b)-standard	Convex	$\mathcal{C}^k$ without boundary	$\mathcal{C}^2$ with boundary
Rate	$n^{-1/b}$	$n^{-1/d}$	$n^{-k/d}$	$n^{-2/(d+1)}$

Table 1.1: Convergence rates for Hausdorff support estimation–Noise free cases.

1.4.4 Noisy cases

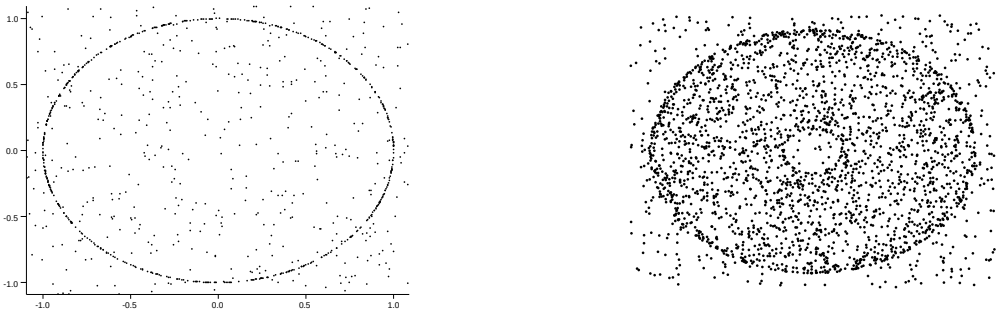
Most of existing results on noisy situations for manifold estimation deal with the case where M is $\mathcal{C}^2$ without boundary. We briefly introduce the different types of noise that are commonly investigated, and give a brief summary of the aforementioned results.

1.4.4.1 Clutter noise

The *clutter noise* assumption introduces some ambient noise. Namely, for a mixing parameter β, the observations X_i's are assumed to have distribution

$$Q \sim \beta P + (1-\beta)\mathcal{U}_{\mathrm{B}(0,R)},$$

where $P \in \mathcal{P}^{2,d,D}_{\tau_{\min},\infty}(f_{\min}, f_{\max})$ and $\mathrm{B}(0,R)$ is the D-dimensional unit ball (that we assume to contain the support of P). With a slight abuse of notation, we denote by $\beta\mathcal{P}^2_{\tau_{\min}} + (1-\beta)\mathcal{U}_{\mathrm{B}(0,R)}$ the set of such distributions.



(a) Circle with outliers: $d = 1$, $D = 2$, $\beta < 1$. (b) Torus with outliers: $d = 2$, $D = 3$, $\beta < 1$.

Figure 1.16: Point clouds $\mathbb{X}_n$ drawn from distributions in $\beta\mathcal{P}^2_{\tau_{\min}} + (1-\beta)\mathcal{U}_{\mathrm{B}(0,R)}$.

Roughly speaking, when facing such a situation, βn sample points are sampled on the manifold, and $(1-\beta)n$ are spurious points lying in the D-dimensional ball $\mathrm{B}(0,R)$ (see

Figure 1.16). To detect those spurious points, the intuition is simple: if x denote a point, T is the tangent space at $\pi_M(x)$ (whenever $\mathrm{d}(x, M) \leq \tau_M$), and h is a bandwidth, the slab $S_x(h)$ may be defined by

$$S_x(h) = \left\{ y \mid \|\pi_T(y - x)\| \leq h \quad \text{and} \quad \|\pi_{T^\perp}(y - x)\| \leq h^2 \right\}.$$

If $\mathrm{d}(x, M) \geq h^2$, the Q-volume of $S_x(h)$ will be roughly h^{2D-d}, whereas it is of order h^d (that is much larger) whenever $x \in M$. This provides a way to detect outliers, based on counting sample points in slabs, that is the core of the following upper bounds.

THEOREM 1.37 :

- *[Genovese et al., 2012a, Theorem 5], [Aamari and Levrard, 2018, Theorem 7]: There exists $\hat{M}$ such that, for n large enough,*

$$\sup_{Q \in \beta \mathcal{P}^2_{\tau_{\min}} + (1-\beta)\mathcal{U}_{\mathrm{B}(0,R)}} \mathbb{E}\mathrm{d}_{\mathrm{H}}(\hat{M}, M) \leq C_{d,\tau_{\min},f_{\min},f_{\max}} \left(\frac{\log(n)}{\beta n} \right)^{\frac{2}{d}}.$$

- *[Genovese et al., 2012a, Theorem 4]:*

$$\inf_{\hat{M}} \sup_{Q \in \beta \mathcal{P}^2_{\tau_{\min}} + (1-\beta)\mathcal{U}_{\mathrm{B}(0,R)}} \mathbb{E}\mathrm{d}_{\mathrm{H}}(\hat{M}, M) \geq c_{d,\tau_{\min},f_{\min},f_{\max}} \left(\frac{1}{\beta n} \right)^{\frac{2}{d}}.$$

The intuition behind Theorem 1.37 is that in the clutter noise model, detecting the $(1 - \beta)n$ outliers is somehow easy enough to get the same convergence rate as in the $\mathcal{C}^2$ case based on the βn signal points. The two mentioned procedures for the upper bound are based on outlier detection based on slabs. The one proposed in Genovese et al. [2012a] consider an empirical fit maximizer over the set of $\mathcal{C}^2$ manifolds, where the fit is defined as the number of points that fall into all possible slabs. The estimator exposed in Aamari and Levrard [2018] relies on an iterative tangent space and denoising procedure: for a subset of $\mathbb{X}_n$, an iteration computes tangent space estimates at subset points and discard points whose corresponding slabs have too small weight, leading to a smaller subset of points. In both cases the ambient dimension D plays no role in the convergence rate. Let us mention however that the prescribed lower bound on n in fact depends on D.

1.4.4.2 Offset noise models

The offset noise models assume that we observe $(Y_1, \dots, Y_n) = \mathbf{Y}_n$ drawn on the offset $M^\sigma = \{x \in \mathbb{R}^D \mid \mathrm{d}(x, M) \leq \sigma\}$. This kind of assumption encompasses the uniform noise model investigated in Aizenbud and Sober [2021], Sober et al. [2021], Maggioni et al. [2016], Aamari et al. [2019], but also structured additive noise models as in Genovese et al. [2012b], Puchkin and Spokoiny [2019]. To simplify, we give results for the uniform noise model, defined by

$$\mathcal{U}^k_{\tau_{\min}}(\sigma) = \left\{ \mathcal{U}(M_\sigma) \quad \text{s.t.} \quad M \in \mathcal{M}^{k,d,D}_{\tau_{\min},\infty} \right\},$$

and the tubular noise model, defined by

$$\mathrm{Tub}^k_{\tau_{\min}}(\sigma) = \Big\{ Y = X + \varepsilon \quad \text{s.t.} \quad X \in \mathcal{P}^{k,d,D}_{\tau_{\min},\infty}(f_{\min}, f_{\max}), \\ (\varepsilon \mid X) \in \mathrm{B}_{(T_X M)^\perp}(\sigma), \quad \mathbb{E}(\varepsilon \mid X) = 0 \Big\},$$

with a slight abuse of notation. Up to our knowledge, these two models are the only ones that account for additive normal noises in the literature. The difference between these two models is the additive noise centering. Indeed, using for instance [Federer, 1959, Theorem 3.1], if $Y \sim Q \in \mathcal{U}_{\tau_{min}}(\sigma)$, then $Y = \pi_M(Y) + (Y - \pi_M(Y)$ can be written as $Y = X + \varepsilon$, with $X \in \mathcal{P}^{k,d,D}_{\tau_{\min},\infty}(f_{\min}, f_{\max})$ and $(\varepsilon \mid X) \in \mathrm{B}_{(T_X M)^\perp}(\sigma)$. However, curvature of M can prevent $\mathbb{E}(\varepsilon \mid X)$ to be null.

Example 1.38. Assume that Y has a uniform distribution over the annulus $C(R-\sigma, R+\sigma) \subset \mathbb{R}^2$. Then $Y = X + \varepsilon$, where X is the uniform distribution on the unit circle, and $(\varepsilon \mid X) \in \mathrm{B}_{(T_X M)^\perp}(\sigma)$. However,

$$\mathbb{E}\left(\varepsilon \mid X\right) = \frac{\sigma^2}{3R} X \neq 0.$$

To sum up, uniform noise may be considered as additive normal and non-centered noise, and conversely tubular noise is noise that has density with respect to the uniform distribution on the offset M^σ. The upper bounds for the offset noise model are summarized below.

THEOREM 1.39 :

Let $Y \sim Q \in \mathcal{U}^k_{\tau_{\min}}(\sigma)$.

- *Maggioni et al. [2016]: For $k = 2$ and $\sigma < (C \log(n)/n)^{\frac{1}{d+2}} < \tau_{\min}$, a local linear patch estimator $\hat{M}$ achieves*

$$\mathbb{E}\left\|\pi_{\hat{M}}(Y) - \pi_M(Y)\right\| \leq C_{d,D,\tau_{\min}} \left(\frac{\log(n)}{n}\right)^{\frac{2}{d+2}}.$$

- *[Aizenbud and Sober, 2021, Theorem 3.3]: For $\sigma < C\tau_{\min}\sqrt{D\log(D)}$, there is an iterative algorithm $\mathcal{A}$ based on $\mathbb{Y}_n$ such that, for any $y \in M^\sigma$ there exists $\hat{p} \in M$ satisfying*

$$\mathbb{E}\left\|\hat{p} - \mathcal{A}(y)\right\| \leq C_{d,k,D,\tau_{\min}} \left(\frac{\log(n)}{n}\right)^{\frac{k}{2k+d}}.$$

Note that the first bound upper bound given in Theorem 1.39 differs from the results exposed in Maggioni et al. [2016]. It can be derived from the proof of [Maggioni et al., 2016, Theorem 6]. The result from Aizenbud and Sober [2021] roughly ensures that $\sup_{x \in \mathcal{A}(M^\sigma)} \mathrm{d}(x, M) \lesssim n^{-\frac{k}{2k+d}}$.
These two results address a slightly different problem from the originate manifold estimation problem. Indeed, both are focused on sending an ambient point onto the manifold. However, it is likely that these two rates can be retrieved for the manifold estimation problem, for instance choosing $\hat{M}$ as the set of projections of a thin enough grid of the ambient ball $\mathrm{B}(0, R)$, or using some polynomial patches around them. The latter approach seems more relevant, since both methods exposed in Aizenbud and Sober [2021], Maggioni et al. [2016] have a tangent space estimation procedure step, combined with a local polynomial fit in Aizenbud and Sober [2021]. The optimality of these rates might be discussed considering the following lower bound.

THEOREM 1.40 : [AAMARI AND LEVRARD, 2019, THEOREM 7]

For a fixed $\tau_{\min}$ and $\mathbf{L}$ large enough, it holds, for all $\sigma < \tau_{\min}$,

$$\inf_{\hat{M}} \sup_{Q \in \mathcal{U}^k_{\tau_{\min}}(\sigma)} \mathbb{E}\left(\mathrm{d_H}(\hat{M}, M)\right) \geq c_{d,k,\tau_{\min}} \left(\frac{\log(n)}{n}\right)^{\frac{k}{k+d}}.$$

Note that the same result for $k = 2$ is to be found in [Genovese et al., 2012b, Theorem 1] Considering Theorems 1.39 and 1.40, we see that there is a gap between the $n^{-k/(2k+d)}$ rate for the upper bound and the $n^{-k/(k+d)}$ rate for the lower bound. To determine which of the two bounds is optimal, if any, is still an open question for $k \geq 3$, up to our knowledge. A similar phenomenon occurs for the tubular noise model, the corresponding upper bounds are given below.

THEOREM 1.41 :

Let $Q \in \mathrm{Tub}^k_{\tau_{\min}}(\sigma)$.

- *Genovese et al. [2012b] For $k = 2$, $(\varepsilon \mid X) \sim \mathcal{U}\left(\mathrm{B}_{(T_X M)^\perp}(\sigma)\right)$ and $\sigma < \tau_{\min}$, there exists $\hat{M}$ such that, for n large enough,*

$$\mathbb{E}\left(\mathrm{d_H}(\hat{M}, M)\right) \leq C_{\sigma,d,D,f_{\min},f_{\max},\tau_{\min}} \left(\frac{\log(n)}{n}\right)^{\frac{2}{d+2}}.$$

- *Puchkin and Spokoiny [2019] For $k = 2$ and $\sigma \lesssim n^{-\frac{2}{3d+8}}$, there exists a Nadaraya-Watson type estimator $\hat{M}$ such that, for n large enough*

$$\mathbb{E}\left(\mathrm{d_H}(\hat{M}, M)\right) \leq C_{f_{\min},f_{\max}} \tau_{\min}^{-1} \left(\frac{D\sigma^2\tau_{\min}^2 \log(n)}{n}\right)^{\frac{2}{d+4}}.$$

The result of Genovese et al. [2012b] is stated for the particular case where the tubular noise is uniform on the orthogonal fiber, whereas the results in Puchkin and Spokoiny [2019] also allows for small tangential noise. Moreover, this $n^{-2/(d+4)}$ convergence rate has been proved optimal in Puchkin and Spokoiny [2019] for $k = 2$ when a small tangential noise is allowed. Up to our knowledge, there is no minimax lower bound for the tubular noise model.

Gathering all these results and remarks, we see that identifying the correct convergence rate for the offset noise model and for the tubular noise model globally remains an open question that depend on the structure of the noise. Both of these models may be written as $Y = X + \varepsilon$, where $\varepsilon \mid X \in \mathrm{B}_{(T_X M)^\perp}(\sigma)$, they differ by the structural assumptions on $\varepsilon \mid X$ (centered or not, has uniform distribution on $\mathrm{B}_{(T_X M)^\perp}(\sigma)$, yields uniform distribution on M^σ). We strongly believe that the key assumption is the (a, b)-standardness of the noise distribution in the normal fibers. To be more precise, we conjecture that the correct convergence rate for "uniform-like" noises on the normal fiber (that is with lower bounded density, encompassing the uniform offset noise model as well as the tubular noise model) is $n^{-\frac{k}{k+d}}$, and that the convergence rate deprecates whenever this orthogonal noise is no more $(a, (D-d))$-standard. This is ongoing work with E. Aamari.

1.4.4.3 Gaussian noise model

As emphasized in the last Section, in the case of centered additive noise, the structure of the noise with respect to the geometry of the underlying manifold plays an important role in the convergence rates. We mention here the results pertaining to a specific instance of unstructured noise model,

$$\mathcal{G}^k_{\tau_{\min}}(\sigma) = \Big\{ Y = X + \varepsilon \quad \text{s.t.} \quad X \in \mathcal{P}^{k,d,D}_{\tau_{\min},\infty}(f_{\min}, f_{\max}), \\ \varepsilon \sim \mathcal{N}(0, \sigma^2), \quad \varepsilon \perp\!\!\!\perp X \Big\},$$

that is the Gaussian noise model. The only results on the convergence rates in such a model are given in Genovese et al. [2012a] for the $\mathcal{C}^2$ case, summarized below.

THEOREM 1.42 :

- *[Genovese et al., 2012a, Theorem 8]: For n large enough,*

$$\inf_{\hat{M}} \sup_{Q \in \mathcal{G}^k_{\tau_{\min}}(\sigma)} \mathbb{E}\left(\mathrm{d_H}(\hat{M}, M)\right) \geq C_{d,D,\tau_{\min},f_{\min},f_{\max},\sigma} \frac{1}{\log n}.$$

- *[Genovese et al., 2012a, Theorem 9]: Let $Q \in \mathcal{G}^k_{\tau_{\min}}(\sigma)$, and $0 < \delta < (1/2)$. Then there exists a deconvolution-based estimator $\hat{M}$ such that, for n large enough,*

$$\mathbb{E}(\mathrm{d_H}(\hat{M}, M)) \leq C_{d,D,\tau_{\min},f_{\min},f_{\max},\sigma,\delta} \left(\frac{1}{\log n}\right)^{\frac{1-\delta}{2}}.$$

Intuitively, Theorem 1.42 assesses that manifold estimation with Gaussian additive noise is at least as difficult as the regression with error in variables problem (Fan and Truong [1993]), from a statistical point of view, as detailed in [Genovese et al., 2012a, Section 6]. The similarity between these two problems is clear, at least locally: from the parametrization described in 1.30, we see that allowing noise in the tangential directions boils down to the observation of both a perturbed regression function (parametrization + noise in the normal direction) and perturbed variables (parameter/projection onto the tangent space + noise in the tangent direction). This justifies in some sense the slow convergence rates in $(\log n)^{-\beta}$ (as in Fan and Truong [1993]).
To give a partial conclusion, we can infer from Theorem 1.37, Theorem 1.39 and Theorem 1.42 that, in the additive noise case, the less the noise is tied to the manifold structure, the slower the convergence rates for Hausdorff estimation will be. The Gaussian noise model may be thought of as a limiting case, that allows too much error in variables for a manifold estimator to converge faster than in $(\log n)^{-1}$.

1.5 Discussion and directions for future research

Tractable manifold estimators for the $\mathcal{C}^k$ case (with boundary).
The local polynomial estimator of M in the $\mathcal{C}^k$ model that is proposed in (1.9) is more a theoretical object than an efficient manifold estimator. We list here some improvements that are in our opinion worth trying. First, from a computational viewpoint, the canonical basis of monomials to approximate a polynomial approximation of the parametrization leads to poor conditioning constants (to relate the square deviation and the coefficients). This fact somehow appears in the constants of Theorem 1.35, where the dependency in k is at least 4^k, due to this conditioning issue. Using a more L_2 dedicated

basis, such as multivariate Lagrange polynomials, could be a way to alleviate this conditioning problem, at the price of the loss of differential quantities estimators that naturally arise from the monomial estimates (such as curvature, see e.g. [Aamari and Levrard, 2019, Section 3.2]). More generally, understanding more closely the influence of the regularity k in the convergence rates would pave the way to adaptivity with respect to k, using standard model selection procedures for instance.
Second, since it consists of a mere union of polynomial patches, this estimator has poor structure. To provide a manifold as estimator, a first step could be to "glue" patches using a kernel approach in the spirit of Puchkin and Spokoiny [2019]. Providing a minimax and topologically correct estimator as for the $\mathcal{C}^2$ case would require more technical work, for instance by replacing the convex hull of d-dimensional simplices of the Tangential Delaunay complex with small polynomial caps in the reconstruction. Such an approach would also require some gluing procedure. Since the topology of such an estimator would still be based on the TDC complex topology, topological correctness is likely to follow.
At last, manifold estimation in the $\mathcal{C}^k$ case with boundary is still an open question. Some insights may be gained in the special case where $D = d + 1$ in the offset noise model: in this case, estimating M via $\hat{M}$ naturally provides an estimator of ∂M^σ via $\{x \mid \mathrm{d}(x, \hat{M}) = \sigma\}$. This yields a $n^{-\frac{k}{k+d}}$ rate for the estimation of the boundary of M^σ, that is in line with the $n^{-\frac{2}{2+(D-1)}}$ rate for boundary estimation of M^σ for the $\mathcal{C}^2$ case. We conjecture that the boundary estimation rate for the $\mathcal{C}^k$ case is indeed $n^{-\frac{k}{k+(d-1)}}$, that is also the correct rate for the manifold estimation problem.

Offset noise models and regression

Informal connection between manifold estimation problem and regression has been established in Genovese et al. [2012a]. Indeed, locally around some point $x \in M$, if T is such that $\angle(T, T_x M) < 1$, then M can be parametrized by $\psi(u)$, where $u \in T$. Adding normal noise then amounts to observe $(U + \varepsilon_1, \psi(U) + \varepsilon_2)$, that is a regression problem with error in variables. Through this lens, the $n^{-\frac{k}{2k+d}}$ rate given in Aizenbud and Sober [2021], Puchkin and Spokoiny [2019] is in line with the convergence rate for non-parametric regression Stone [1982]. However, the $n^{-\frac{2}{2+d}}$ rate obtained in Genovese et al. [2012a] suggests that the manifold estimation problem with tubular noise is somehow easier than the regression problem in full generality. We conjecture that this is due to the particular structure of the noise: in a regression setting with bounded noise $\varepsilon \in [-\sigma, \sigma]$, faster convergence rates may be obtained with a local extrema based estimator, of the type $\hat{f}(x) = \min_{\|X_i - x\| \leq h} Y_i + \sigma$, provided that $\mathbb{P}(\varepsilon \in [-\sigma, -\sigma + \delta])$ is large enough with respect to δ. This intuition might be adapted to the tubular and offset noise models for manifold estimation: we conjecture that the $n^{-\frac{k}{k+d}}$ rate is achievable whenever the noise distribution spreads enough weight around its extrema, the centering assumption playing no role in this case.

Byproducts of manifold estimation.

The manifold estimation procedures we described allows to infer some manifold parameters, based on $\hat{M}$ or its construction. For instance, the local polynomials estimation that is carried out in Section 1.4.3 provides estimates of tangent spaces and second fundamental form (see also [Aamari and Levrard, 2019, Section 3.1,3.2]). These estimates can be used to build estimators of the mean curvature Aamari and Levrard [2019], they can be further exploited to build estimators of the reach based on the heuristics exposed in Aamari et al. [2019], Berenfeld et al. [2020].
Other quantities such as the volume of the manifold or that of its boundary may be estimated using estimators $\hat{M}$, using for instance the local decomposition of the volume of M^r for $r \to 0$ given in Federer [1959]. Note that this decomposition (based on

curvature measures) heavily depends on the existence of a boundary, so that the procedure described in Section 1.4.2 is of particular interest in this case.

Chapter 2

Coresets for distance to support estimation

This chapter exposes and motivates the robust quantization scheme we conceived in Brécheteau and Levrard [2020] in a compact set estimation framework. The key notion to connect compact set estimation and quantization is the notion of coreset: for a target precision ε we intend to build ε-close estimators of d_K based on a small number k of points, of the type $\mathrm{d}_{c_1,\dots,c_k}$ for instance, where the c_j's are points in $\mathbb{R}^D$.

First, we motivate the design of k-points approximations at a macroscopic scale (of order τ_K), by introducing some topological features of K that can be retrieved from such an approximation, such as its persistence diagram.

Next, a short overview of selected quantization results takes place, that explains the expected benefits of the quantization approach in a noisy compact set estimation setting, as well as its limitation. To summarize, the quantization approach intends to minimize $P(du)\|\mathrm{d}_{c_1,\dots,c_k}(u) - \mathrm{d}_K(u)\|^2$, rather than $\|\mathrm{d}_{c_1,\dots,c_k} - \mathrm{d}_K\|_\infty$. Adopting such a L_2-like criterion renders the estimation less sensitive to additive noise. However, providing guarantees on the Hausdorff precision of $\mathrm{d}_{c_1,\dots,c_k}$ in noisy settings needs some adaptation of this quantization principle.

The k-PDTM (k-power distance to measure) we propose consists in replacing d_K by a robust distance approximation (the distance to measure). This heuristic is described in Section 2.3, along with theoretical guarantees on the sample-based k-PDTM in terms of Hausdorff precision. These results guarantee that topological inference based on the k-PDTM is a sensible strategy in additive noise cases. This statement is alleged by numerical experimentation in a persistence diagram estimation framework.

2.1 Motivation and noise-free case

2.1.1 Topological inference from distance to support

The previous chapter was devoted to get, under various regularity settings, optimal estimators of d_K. Besides from the compact reconstruction problem, such distance estimators may be used to retrieve some topological properties of the compact K from their sublevel sets. In other words, for a distance estimator $\hat{\mathrm{d}}$, we may infer some topological properties of K from $\hat{d}^{-1}[0,r]$, for a suitable choice of r and provided that $\hat{\mathrm{d}}$ looks like a distance function in the following sense.

Definition 2.1 :

A non-negative function $\phi : \mathbb{R}^D \to \mathbb{R}^+$ is *distance-like* if there exists $I \subset \mathbb{R}$ and $\{(c_i, \omega_i)\}_{i \in I} \in (\mathbb{R}^D \times \mathbb{R}^+)^I$ such that

$$\phi^2(x) = \inf_{i \in I} \|x - c_i\|^2 + \omega_i^2,$$

for all $x \in \mathbb{R}^D$.

Note that Definition 2.1 differs from the original definition of distance-like functions given by [Boissonnat et al., 2018, Definition 10.1]. However, [Chazal et al., 2011, Proposition 3.1] ensures that these definitions are equivalent. In particular, a distance-like function in the sense of Definition 2.1 is 1-Lipschitz, 1-semiconcave and proper.
Distance-like functions are roughly the set of power-distance functions. In particular, for any compact $K' \subset \mathbb{R}^D$, $\mathrm{d}_{K'}$ is distance-like, so that for instance $\mathrm{d}_{\mathbb{X}_n}$ (distance to the sample) is distance-like. The interesting point is that, for a distance-like function $\hat{\mathrm{d}}$, L_∞-type bounds for $\hat{\mathrm{d}} - \mathrm{d}_K$ may be translated into topological inference results.

Theorem 2.2 : [Chazal et al., 2011, Theorem 4.6]

Let $\hat{\mathrm{d}}$ be a distance-like function such that

$$\|\hat{\mathrm{d}} - \mathrm{d}_K\|_\infty \leq \varepsilon.$$

If $\varepsilon \leq \frac{\tau_K}{9}$ and $4\varepsilon < r < \tau_K - 3\varepsilon$, then K and $\hat{\mathrm{d}}^{-1}([0, r])$ are homotopy equivalent.

In a nutshell, homotopy equivalence ensures that $\hat{\mathrm{d}}^{-1}([0, r])$ continuously retracts on K. In particular, homotopy equivalence ensures that the singular homology groups of K and $\hat{\mathrm{d}}^{-1}([0, r])$ are isomorphic (see e.g., [Munkres, 1984, Theorem 30.8]). To give an intuition, the k-th homology group of K is spanned by the k-dimensional holes of K. For instance, if K has a single connected component, then the 0-th homology group of K is isomorphic to $\mathbb{F}_2$. If K has two 1-dimensional holes, that is portions of K in which closed loops cannot continuously retract on a single point, then the first homology group of K is isomorphic to $\mathbb{F}_2^2$, etc. For $k \geq 0$, the number of k-dimensional holes of K is the k-th *Betti number*, that is just the dimension of the k-th homology group. For a general introduction to simplicial homology the interested reader is referred to Munkres [1984], Hatcher [2002].
According to Theorem 2.2, it is enough to provide a $\tau_K/9$-close estimation of K in terms of Hausdorff distance to recover its homology groups (via the homology groups of the sublevel sets of the corresponding distance $\hat{\mathrm{d}}$). For instance, if $\hat{\mathrm{d}} = \mathrm{d}_{\mathbb{X}_n}$, Theorem 2.2 ensures that the Devroye-Wise estimator $\mathrm{B}(\mathbb{X}_n, r)$ of K has the correct singular homology structure, provided n is large enough (depending on regularity assumptions on K) and that the radius r of the balls is suitably chosen.

2.1.1.1 Persistence diagrams

To tackle the tough question of the selection of r, or simply to track multiscale information, the concept of persistent homology has been introduced in Edelsbrunner et al. [2002]. It consists in describing the evolution of the homology of the sublevel sets of a function f. Namely, for $f : \mathbb{R}^D \mapsto \mathbb{R}$, not only the homology groups of $f^{-1}((-\infty, r])$ are described, but also the morphisms between homology groups of $f^{-1}((-\infty, r])$ and

$f^{-1}((-\infty, s])$ the inclusion map induces, allowing to keep tracks of apparitions and collapses of topological features.
Persistent homology can be encoded via persistence diagrams. A persistence diagram is a multiset of points (b, d). Each point (b, d) is associated to one topological feature (that accounts for one dimension of the corresponding homology group: a connected component, a hole, a void, etc.) that appears when $r = b$ (its birth time) and disappears when $r = d$ (its death time). To give an intuition, Figure 2.1 below exposes the persistence diagram of d_K, where K is the (unbalanced) infinity symbol. Slightly anticipating, this persistence diagram is well-defined.

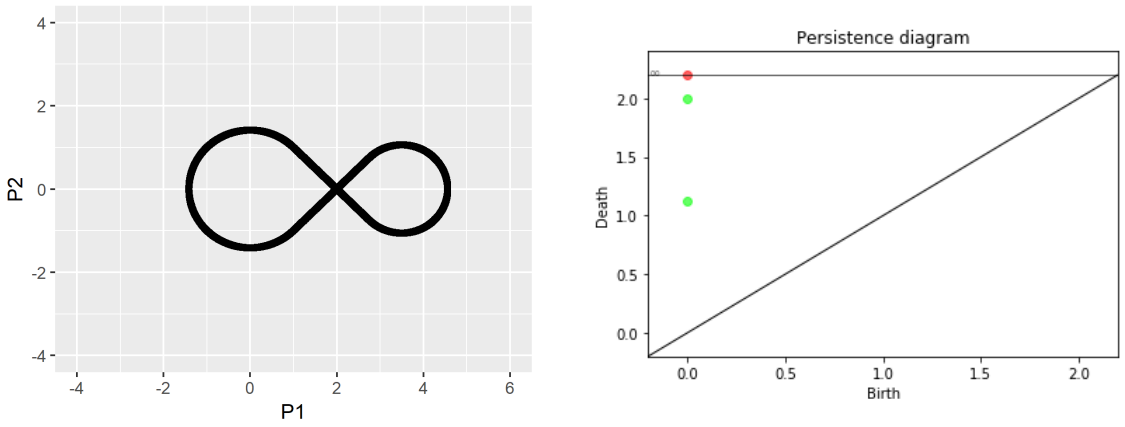


Figure 2.1: Persistence diagram for the infinity symbol: The red point accounts for the (infinitely persistent) connected component. The two green points represent the two loops (one-dimensional holes).

Note that points of a persistence diagrams are located in the one eight of plane defined $\{(x, y) \in (\mathbb{R}^+)^2 \mid y > x\}$. If Δ denotes the diagonal $\{(x, y) \in (\mathbb{R}^+)^2 \mid y = x\}$, and (b, d) is a points of a persistence diagram corresponding to a topological feature, the distance of (b, d) to Δ may be interpreted as the persistence of the topological feature, that seizes how long this feature is represented in the filtration. For technical reasons, the persistence diagrams include the diagonal Δ, with infinite multiplicity. For a general introduction to persistence diagrams the interested reader is referred to [Boissonnat et al., 2018, Section 11.5]. Persistence diagrams are well-defined for functions f that satisfy the *q-tameness* regularity assumption (see [Boissonnat et al., 2018, Definition 11.20]). To spare the reader the general definition of q-tameness, the following result ensures that distance functions are q-tame.

THEOREM 2.3 : [BOISSONNAT ET AL., 2018, THEOREM 11.28]

> *Let $K \subset \mathbb{R}^D$ be a compact set. Then d_K is q-tame.*

In particular, $\mathrm{d}_{\mathbb{X}_n}$ is also q-tame, so that persistence diagrams for the distance to a finite point set is also well-defined. It follows from the proof of Theorem 2.3 that *weighted* distances to a finite set of points, that is functions of the form $x \mapsto \sqrt{\min_{j=1,\dots,k} \|x - c_j\|^2 + \omega^2(c_j)}$ are also q-tame. Slightly anticipating, since all considered functions for which persistence diagrams will be build are weighted distance to finite sets of points or distance to compact sets, it is implicitly assumed that the q-tameness assumption is satisfied (and thus the persistence diagram is well defined). For such a function, we denote by $D(f)$ its persistence diagram.
To assess proximity between persistence diagrams, a distance between diagrams is

needed. The most common choice is the so-called *bottleneck distance*, defined below.

Definition 2.4 :

Let D_1 and D_2 be persistence diagrams. A pairing between D_1 and D_2 is a bijection between the multisets of points D_1 and D_2 (recall that they both contain the diagonal with infinite multiplicity).

The bottleneck distance is then defined by

$$\mathrm{d}_B(D_1, D_2) = \inf_{\text{pairings } \gamma} \sup_{p \in D_1} \|p - \gamma(p)\|_\infty ,$$

with the convention $\|(b_p, +\infty) - (b_q, +\infty)\|_\infty = |b_p - b_q|$, to account for the connected component with infinite persistence.

Intuitively speaking, the bottleneck distance may be thought of as a L_∞ transportation distance between diagrams, where each point can be mapped to the diagonal. See Figure 2.2 below to get an intuition on pairings and bottleneck distance. If persistence diagrams

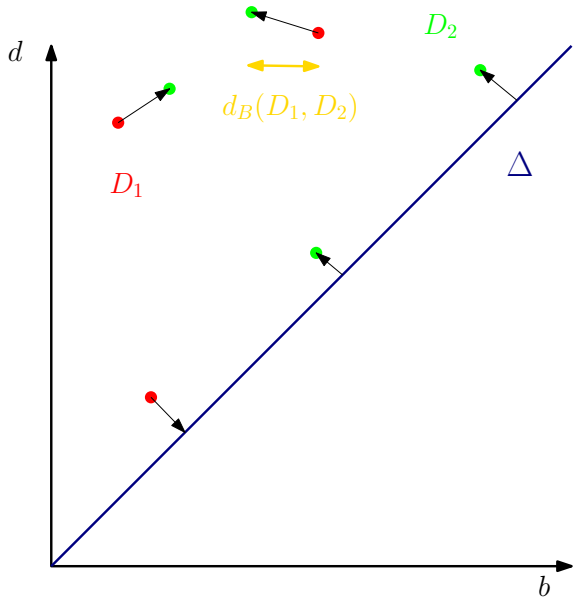


Figure 2.2: An optimal pairing between D_1 (red) and D_2 (green). The bottleneck distance is depicted in gold

are built from the sublevel sets of tame functions f and g, then the following stability results connects bottleneck distance between $D(f)$ and $D(g)$ to the sup-norm deviation between f and g.

Theorem 2.5 : Cohen-Steiner et al. [2007]

If f and g are q-tame functions, then

$$\mathrm{d}_B(D(f), D(g)) \leq \|f - g\|_\infty .$$

In the particular distance estimation case, Theorem 2.5 ensures that the persistence diagrams $D(\hat{\mathrm{d}})$ and $D(\mathrm{d}_K)$ are close whenever $\|\hat{\mathrm{d}} - \mathrm{d}_K\|_\infty$ is small, that is the lifetimes $d - b$ of the topological features will be similar. In this case, let $\delta_{\min}$ denote the smallest distance of points in $D(\mathrm{d}_K)$ (the diagonal being excluded) to Δ, and $D(\hat{\mathrm{d}})_{\geq \varepsilon}$ the set of

points in $D(\hat{\mathrm{d}})$ that have persistence larger than ε. Then $D(\mathrm{d}_K)$ and $D(\hat{\mathrm{d}})_{\geq\varepsilon}$ have the same structure whenever $\varepsilon < \delta_{\min}/2$. In particular they have the same number of points (see Figure 2.3 below).

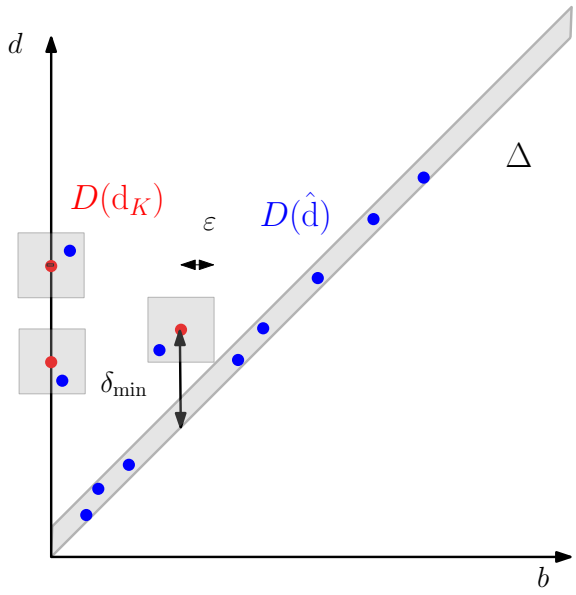


Figure 2.3: Points of $D(\hat{\mathrm{d}})$ farther than ε to the diagonal account for "true" topological features of K.

To conclude this section, let us emphasize that topological inference based on the sublevel sets of a distance estimator $\hat{\mathrm{d}}$ needs a prescribed precision in terms of $\|\hat{\mathrm{d}} - \mathrm{d}_K\|_\infty$ to provide relevant topological descriptors. This precision threshold can be $\tau_K/9$ for homology inference, or $\delta_{\min}/2$ for persistence diagram inference.

2.1.2 Noise-free case and coresets

As exposed above, a standard way to estimate the topology of K is to compute the singular homology groups of $\mathrm{B}(\mathbb{X}_n, r)$, for a suitable choice of r. In practice, the singular homology of $\mathrm{B}(\mathbb{X}_n, r)$ is retrieved via the computation of the simplicial homology of the Čech complex $\check{\mathrm{C}}\mathrm{ech}(\mathbb{X}_n, r)$ defined by

$$\{X_{i_0}, \ldots, X_{i_k}\} \in \check{\mathrm{C}}\mathrm{ech}(\mathbb{X}_n, r) \quad \Leftrightarrow \quad \bigcap_{j=0,\ldots,k} \mathrm{B}(X_{i_j}, r) \neq \emptyset.$$

Indeed, the Nerve Theorem [Chazal et al., 2011, Theorem 2.8] ensures that $\mathrm{B}(\mathbb{X}_n, r)$ and $\check{\mathrm{C}}\mathrm{ech}(\mathbb{X}_n, r)$ are homotopy equivalent, and, for the simplicial complex $\check{\mathrm{C}}\mathrm{ech}(\mathbb{X}_n, r)$, the singular homology groups are isomorphic to simplicial homology groups. For more details on simplicial complexes and their use in topological inference the interested reader is referred to [Boissonnat et al., 2018, Chapter 2 and 11]. As well, computing the persistence diagram of $\mathrm{d}_{\mathbb{X}_n}$ may be carried out by computing the persistence diagram of $\check{\mathrm{C}}\mathrm{ech}(\mathbb{X}_n, r)$, for increasing $r \in [0, +\infty)$.
From a computational viewpoint, the Čech complex for a radius r may be computed with $O(n \times (d+1)!)$ operations in expectation (see, e.g., [Edelsbrunner and Harer, 2010, Section III.2]), that may become intractable for large dimensions. The Vietoris-Rips complex $\mathrm{Rips}(\mathbb{X}_n, r)$, defined by

$$\{X_{i_0}, \ldots, X_{i_k}\} \in \mathrm{Rips}(\mathbb{X}_n, r) \quad \Leftrightarrow \quad \forall j \neq \ell \quad \mathrm{B}(X_{i_j}, r) \cap \mathrm{B}(X_{i_\ell}, r) \neq \emptyset,$$

is often preferred to compute simplicial homology groups. Indeed, it is immediate that

$$\mathrm{Rips}(\mathbb{X}_n, r) \subset \check{\mathrm{C}}\mathrm{ech}(\mathbb{X}_n, r) \subset \mathrm{Rips}(\mathbb{X}_n, 2r),$$

so that the homology groups of K may be retrieved from those of $\mathrm{Rips}(\mathbb{X}_n, r)$ for $2\varepsilon \leq r < \frac{1}{4}(\tau_K - \varepsilon)$, whenever $\|\mathrm{d}_{\mathbb{X}_n} - \mathrm{d}_K\|_\infty \leq \varepsilon < \tau_K/9$ (see, e.g., [Chazal and Oudot, 2008, Theorem 3.6]). In this case, only pairwise distances between sample points are needed to compute the Rips complex, that is $O(dn^2)$ operations. Computing the homology of this complex may be carried out using Zomorodian and Carlsson [2005]. Whatever the simplicial complex (and algorithm) used, it is immediate that if $C \subset \mathbb{X}_n$ is a $\tau_K/18$-covering of $\mathbb{X}_n$ and $\mathbb{X}_n$ is a $\tau_K/18$-covering of K, then

$$\|\mathrm{d}_C - \mathrm{d}_K\|_\infty \leq \frac{\tau_K}{18} + \|\mathrm{d}_{\mathbb{X}_n} - \mathrm{d}_K\|_\infty \leq \frac{\tau_K}{9},$$

so that the correct homology groups might be computed from $\mathrm{Rips}(C, r)$, at a lower computational cost (depending on the cardinality of C). In the noise-free case, given a precision level (for instance $\tau_K/18$), such a sub-sample may be extracted from $\mathbb{X}_n$ via the Farthest Point Sampling algorithm, and may have a much smaller cardinality than n, depending on the sampling assumptions. For instance, if the source distribution P is (a, b)-standard, then using Theorem1.4 yields $|C| \lesssim \tau_K^{-b}$, with high probability. Thus, in the (a, b)-standard noise-free case, only $O(\tau_K^{-b})$ points are needed to compute a correct topology estimator. Those points form a *coreset* for d_K at precision $\tau_K/9$. This application in topological inference is to be kept in mind to justify the introduction of dedicated methods in the sections to follow.
Once motivated, let us now define the notion of *coreset* that we will use throughout this chapter.

Definition 2.6 : Coresets

Let $\phi^2 : x \mapsto \inf_{i\in I} \|x - c_i\|^2 + \omega_i^2$ be a distance-like function, and $\varepsilon > 0$, $k \in \mathbb{N}^*$. A k-points *coreset* for ϕ with precision ε is a subset of $(\mathbb{R}^D \times \mathbb{R}^+)^k$, $\{\tilde{c}_j, \tilde{\omega}_j\}_{j=1,\dots,k}$, such that

$$\left\|\phi - \tilde{\phi}\right\|_\infty \leq \varepsilon,$$

where

$$\tilde{\phi} : \begin{cases} \mathbb{R}^d & \to & \mathbb{R}^+ \\ x & \mapsto & \min_{j=1,\dots,k} \sqrt{\|x - \tilde{c}_j\|^2 + \omega_j^2} := \min_{j=1,\dots,k} \tilde{\phi}_j(x) \end{cases}.$$

This definition of a coreset is tied to the topological inference framework: since $\tilde{\phi}$ is distance-like, homology of the sublevel sets of ϕ are related to those of $\tilde{\phi}$, using [Chazal et al., 2011, Proposition 4.3] for instance. We might figure out cases where the function to approximate is not distance-like, or where the requirements on the approximation functions $\tilde{\phi}_j$ are different, for instance allowing weighted distances to base structures such as ellipsoids rather than points. Up to our knowledge, the formal definition of a coreset may vary depending on the field and purpose of such an approximation. Most of existing references can be found in the computer science literature. Definition 2.6 is thus an attempt to formalize the notion of k-points distance approximation that fits our purpose, and should not be taken as a canonical definition.
In the simple case where $\phi = \mathrm{d}_K$ for some compact set K, if $\mathbb{X}_n$ is an n-sample from an (a, b)-standard measure on K, then the output of the Farthest Point Sampling Algorithm with parameter $\varepsilon/2$ provides a ε-coreset for d_K whenever $\mathrm{d_H}(\mathbb{X}_n, K) \leq \varepsilon/2$, by letting $\tilde{\phi}_j = \mathrm{d}(., c_j)$, where $c_j \in C$ (and C is the sub-sample). With a slight abuse of terminology, the set $C = \{\tilde{c}_1, \dots, \tilde{c}_k\}$ is also called a k-points coreset (when the choice of weights $\tilde{\omega}_j$ is clear). The general definition of a coreset also allows other distance-like function to be approximated, that will be of key interest in Section 2.3.

Our aim will to build coresets for distances, based on a sample. In other words, for a precision parameter ε, we intend to build a ε-approximation of f with k simple distances, based on sample. A special interest will be paid on the smallest size of a such a ε-coreset, denoted by $k(\varepsilon)$. Following up Theorem 1.4, whenever P is (a, b)-standard, it holds $k(\varepsilon) \lesssim \varepsilon^{-b}$, and a ε-coreset based on $\mathbb{X}_n$ mat be computed using the Farthest Point Sampling Algorithm on $\mathbb{X}_n$. According to Proposition 1.5, this ε^{-b} asymptotic rate is tight.
Note that the results of Section 1.4 might be adapted as well in terms of "generalized" coresets. Indeed, in the $\mathcal{C}^2$ manifold case without boundary, following Section 1.4.3, if $\{c_1, \ldots, c_k\}$ is the output of the Farthest Point Sampling algorithm with parameter ε, and if $T_1, \ldots, T_k$ are the corresponding estimation of tangent spaces, then defining $\tilde{\phi}_j = \mathrm{d}_{\hat{P}_j}$, where

$$\hat{P}_j = B_{T_j}(c_j, 7\varepsilon/8),$$

leads to

$$\left\| \tilde{\phi} - d_M \right\|_\infty \lesssim \varepsilon^2,$$

with high probability, for n large enough. In this framework, it holds $k(\varepsilon) \lesssim \varepsilon^{-d/2}$. Note however that such an approximation requires the additional input of (estimated) tangent spaces, and falls slightly beyond the scope of Definition 2.6. Though the topological correctness result given by Theorem 1.19 remains valid, the additional computation of tangent spaces and the computational cost of retrieving the homology of the sublevel sets of $\tilde{\phi}$ (that does not resort to Rips complexes in this case) might deprecate the practical interest of such an approximation. Such an approach is also developed in Maggioni et al. [2016]. As well, enabling additional inputs such as higher order derivatives, normal to boundary, etc. would provide improved rates for k-distances approximation, that are straightforward..
To sum up, in the noise-free case, coresets may be easily build using the Farthest Point Sampling algorithm, and further processing as described in Section 1.4 can readily apply (providing more complex distances). Building coresets in noisy cases is more challenging, as exposed in the following section.

2.2 Noisy cases and quantization

From now on we assume that the sample $\mathbb{X}_n$ is drawn from Q, a perturbation of the noise-free distribution P that has support K. In full generality, allowing additive noise ensures that a ε-coreset for d_K would have cardinality larger than $Ce^{-1/\varepsilon}$, according to Theorem 1.42. However, the latter bound is purely asymptotic, so that building a $\tau_K/9$-coreset for d_K with reasonable cardinality might be attempted, whenever $\tau_K > 0$. Note that this would be enough to carry further topological inference, as described below Theorem 2.2. In other words, for a fixed desired precision ε, it is likely that, for a bounded level of noise, reasonable coresets may be conceived.
This regime is different from the framework of Theorem 1.42, where the noise level is fixed, and ε is allowed to be arbitrarily small. It falls under the intuition that a key quantity to determine whether such an approximation is feasible will be the level of noise, that we define by $\sigma = W_2(P, Q)$, the 2-Wasserstein distance between the noisy and the noise-free distribution. Note that other distances between P and Q are possible, however the 2-Wasserstein distance allows to readily apply some results from the quantization theory, as detailed below.
Since Q is allowed to have an unbounded support, procedures that rely on gridding the sample point $\mathbb{X}_n$ such as Farthest Point Sampling are likely to fail. Such procedures are

devoted to minimize $\|\mathrm{d}_K - \mathrm{d}_C\|_\infty$, via minimizing $\|\mathrm{d}_{\mathbb{X}_n} - \mathrm{d}_C\|_\infty$, where $C = \{c_1, \ldots, c_k\}$ is a k-points coreset. To prevent candidate points to grid the whole ambient space, a local mean approach could be preferred. To do so, a suitable strategy is to minimize

$$P(dx)\left|\mathrm{d}_K^2(x) - \mathrm{d}_C^2(x)\right| = P(dx)\mathrm{d}_C^2(x) = P(dx)\min_{j=1,\ldots,k}\|x - c_j\|^2 := R_P(C),$$

rather than $R_{P,\infty}(C) = \|d_K - \mathrm{d}_C\|_\infty$. Minimizing such a criterion is the main purpose of the quantization field. To cope with the quantization terminology, candidate elements $c_1, \ldots, c_k$ of $\mathbb{R}^D$ will be called *code points*, and $\mathbf{c} = (c_1, \ldots, c_k)$ will be called a *codebook*. The only difference with a k-points coreset is the specification of an order between code points, that allows to use some Euclidean properties of $(\mathbb{R}^D)^k$, for instance in Levrard [2013, 2015]. Now, for $\mathbf{c} \in (\mathbb{R}^D)^k$, $R(\mathbf{c})$ is straightforwardly defined by

$$R_P(\mathbf{c}) = R_P(C),$$

where $C = \{c_1, \ldots, c_k\}$. With a slight abuse of notation, for $\mathbf{c} \in (\mathbb{R}^D)^k$, we let $\mathrm{d}_{\mathbf{c}} = \mathrm{d}_C$. Contrary to the targeted L_∞ criterion, this L_2 risk is adapted to perturbations expressed in terms of W_2 distance, as exposed below.

Proposition 2.7 :

Let Q be such that $W_2(P, Q) \leq \sigma$. Then, for $\mathbf{c} \in (\mathbb{R}^D)^k$, we have

$$\left|\sqrt{R_P(\mathbf{c})} - \sqrt{R_Q(\mathbf{c})}\right| \leq \sigma. \tag{2.1}$$

As a consequence, if

$$\mathbf{c}_Q^* \in \arg\min_{\mathbf{c}\in(\mathbb{R}^D)^k} R_Q(\mathbf{c}),$$

then, we have

$$R_P(\mathbf{c}_Q^*) \leq 4\min_{\mathbf{c}\in(\mathbb{R}^D)^k} R_P(\mathbf{c}) + 6\sigma^2.$$

Before turning to the short proof of Proposition 2.7, note that the existence of such a $\mathbf{c}_Q^*$ follows from the assumptions $\mathrm{Supp}(P) = K$ and $W_2(P, Q) \leq \sigma$, according to [Fischer, 2010, Theorem 3.1] or [Graf and Luschgy, 2000, Theorem 4.12]. In what follows, the existence of such an optimal codebook will be implicitly assumed.

Proof of Proposition 2.7. Let $\mathbf{c} \in (\mathbb{R}^D)^k$. To prove (2.1), simply note that $\sqrt{R_P(\mathbf{c})}$ (resp. $\sqrt{R_Q(\mathbf{c})}$) is the Wasserstein distance between P (resp. Q) and $\mathcal{P}_{\mathbf{c}}$, that is the set of probability distributions supported by $\mathbf{c}$. The result then follows from a basic triangle inequality.
Next, if $\mathbf{c}_Q^* \in \arg\min_{\mathbf{c}\in(\mathbb{R}^D)^k} R_Q(\mathbf{c})$ and $\mathbf{c}_P^* \in \arg\min_{\mathbf{c}\in(\mathbb{R}^D)^k} R_P(\mathbf{c})$, we have

$$\begin{aligned} R_P(\mathbf{c}_Q^*) &\leq 2R_Q(\mathbf{c}_Q^*) + 2\sigma^2 \\ &\leq 2R_Q(\mathbf{c}_P^*) + 2\sigma^2 \\ &\leq 4R_P(\mathbf{c}_P^*) + 6\sigma^2. \end{aligned}$$

□

Thus, minimizing $Q(dx)\mathrm{d}_{\mathbf{c}}^2(x)$ allows to approximate $\min_{\mathbf{c}} P(dx)\mathrm{d}_{\mathbf{c}}^2(x)$ for reasonable σ's. Next, bounds on $P(dx)\mathrm{d}_{\mathbf{c}}^2(x)$ may in turn be converted into L_∞-type bounds whenever P

is (a,b)-standard. Interestingly, in this case $\min_{\mathbf{c}} P(dx)\mathrm{d}_{\mathbf{c}}^2(x)$ can also be controlled, as exposed below.

PROPOSITION 2.8 :

Assume that P is (a,b)-standard at scale r_0, with compact support K. Then we have:

(i) For all $k \geq \frac{r_0^{-b}}{a}$,

$$\min_{\mathbf{c}\in(\mathbb{R}^D)^k} P(dx) \min_{j=1,\dots,k} \|x - c_j\|^2 \leq 4(ak)^{-\frac{2}{b}};$$

(ii) For any $\mathbf{c} \in (\mathbb{R}^D)^k$,

$$\|\mathrm{d}_{\mathbf{c}} - \mathrm{d}_K\|_{\infty,K} := \max_{x\in K} |\mathrm{d}_{\mathbf{c}}(x) - \mathrm{d}_K(x)| \leq 2\left[\left(\frac{R_P(\mathbf{c})}{a}\right)^{\frac{1}{b+2}} \vee \sqrt{R_P(\mathbf{c})}\left(1 \vee \frac{1}{\sqrt{ar_0^b}}\right)\right],$$

where we recall that $R_P(\mathbf{c}) = P(dx)\mathrm{d}_{\mathbf{c}}^2(x)$.

Proof of Proposition 2.8. For $\varepsilon > 0$, denote by $\mathcal{N}(\varepsilon)$ the ε-covering number of K, that is the smallest number of closed balls with radius ε that cover K, and by $\mathrm{Pack}(\varepsilon)$ the ε-packing number of K, that is the largest cardinality of a subset of K that consists of ε-separated points. Let us quote a classical inequality relating $\mathrm{Pack}(\varepsilon)$ and $\mathcal{N}(\varepsilon)$:

$$\mathrm{Pack}(\varepsilon) \geq N(\varepsilon) \geq \mathrm{Pack}(2\varepsilon).$$

If P is (a,b)-standard at scale r_0 and $k = \mathrm{Pack}(2\varepsilon)$, denote by $C = \{c_1, \dots, c_k\}$ such a maximal 2ε-net. It holds

$$P\left(\bigcup_{j=1,\dots,k} \bar{B}(c_j, \varepsilon)\right) \leq P(K) \leq 1,$$

so that, for $\varepsilon \leq r_0$,

$$\mathrm{Pack}(2\varepsilon)(a\varepsilon^b \wedge 1) \leq 1.$$

This leads to

$$\mathcal{N}(\varepsilon) \leq \frac{2^b}{a}\varepsilon^{-b} \vee 1,$$

whenever $\varepsilon \leq 2r_0$, hence, for all $\varepsilon > 0$,

$$\mathcal{N}(\varepsilon) \leq \frac{2^b}{a}\varepsilon^{-b} \vee \frac{r_0^{-b}}{a} \vee 1.$$

Now, if $k \in \mathbb{N}^*$ is such that $k \geq \frac{r_0^{-b}}{a}$, then there exists $\mathbf{c}$ a ε_k-covering of K with k points, with

$$\varepsilon_k = 2(ak)^{-\frac{1}{b}},$$

so that

$$\min_{\mathbf{c}} P(dx)\mathrm{d}_{\mathbf{c}}^2(x) \leq R_P(\mathbf{c}) \leq \varepsilon_k^2 = 4(ak)^{-\frac{2}{b}},$$

which proves (i).
To prove (ii), let $\mathbf{c} \in (\mathbb{R}^D)^k$, $\Delta_{\infty,K} = \max_{x \in K} |\mathrm{d}_{\mathbf{c}}(x) - \mathrm{d}_K(x)|$, and let $y \in K$ be such that $|\mathrm{d}_{\mathbf{c}}(y) - \mathrm{d}_K(y)| = \Delta_{\infty,K}$. Then

$$\bar{\mathrm{B}}\left(y, \frac{\Delta_{\infty,K}}{2}\right) \cap K \subset \left\{z \in K \mid |\mathrm{d}_{\mathbf{c}}(z) - \mathrm{d}_K(z)| \geq \frac{\Delta_{\infty,K}}{2}\right\}.$$

Markov's inequality entails that

$$\begin{aligned} R_P(\mathbf{c}) &\geq \frac{\Delta_{\infty,K}^2}{4} P\left(\bar{B}\left(z, \frac{\Delta_{\infty,K}}{2}\right)\right) \\ &\geq \frac{\Delta_{\infty,K}^2}{4}\left(a\frac{\Delta_{\infty,K}^b}{2^b} \wedge a r_0^b \wedge 1\right), \end{aligned}$$

so that

$$\Delta_{\infty,K} \leq 2\left(\frac{R_P(\mathbf{c})}{a}\right)^{\frac{1}{b+2}} \vee 2\left(\frac{R_P(\mathbf{c})}{a r_0^b}\right)^{\frac{1}{2}} \vee 2R_P(\mathbf{c})^{\frac{1}{2}}.$$

□

Proposition 2.8, (ii) allows to convert bounds on $R_P(\mathbf{c})$ onto bounds on the covering radius of K with base points $\mathbf{c}$. Combined with Proposition 2.7 and Proposition 2.8, (i), it ensures that, if $W_2(P, Q) \leq \sigma$ and $\mathbf{c}_Q^*$ is an optimal codebook for Q,

$$\left\|\mathrm{d}_{\mathbf{c}_Q^*} - \mathrm{d}_K\right\|_{\infty,K} \leq C\left(a^{-\frac{1}{b}} k^{-\frac{2}{b(b+2)}} + a^{-\frac{1}{b+2}} \sigma^{\frac{2}{b+2}}\right),$$

whenever P is (a, b)-standard. Turning to the topological inference problem described in Section 2.1, $\left\|\mathrm{d}_{\mathbf{c}_Q^*} - \mathrm{d}_K\right\|_{\infty,K}$ can be smaller than the threshold of Theorem 2.2, $\tau_K/9$, whenever $\sigma \lesssim \tau_K^{1+\frac{b}{2}}$ and k is large enough. Therefore, building coresets with reasonable cardinality seems at first sight possible, in the case where $\sigma \lesssim \tau_K^{1+\frac{b}{2}}$, using quantization related tools.
Besides, in the noise-free case, Proposition 2.8 ensures that optimal codebooks for P yield $O(k^{-\frac{2}{b(b+2)}})$ coverings of the support K. With slightly more requirements on K, this rate can be improved to reach the $k^{-\frac{1}{b}}$ covering rate that a k-points optimal covering induces in this case.

THEOREM 2.9 : FROM [CLARKSON, 2006, THEOREM 5.4]

> *Assume that P is (a, b)-standard with scale r_0, with a connected support K. Then, for k large enough, if $\mathbf{c}^*$ denotes an optimal k-points codebook for P, it holds*
>
> $$\|\mathrm{d}_{\mathbf{c}^*} - \mathrm{d}_K\|_{\infty,K} \leq C(ak)^{-\frac{1}{b}}.$$

The original result from Clarkson [2006] is valid under slightly stronger conditions than (a, b)-standardness. However its proof easily adapts as follows.

Proof of Theorem 2.9. We follow the proof of [Clarkson, 2006, Theorem 5.4]. Let $\mathbf{c}^*$ denote an optimal k-points codebook for P, and denote, for $j = 1, \ldots, k$, by $V_j(\mathbf{c}^*)$ the Voronoi cell of c_j^*. According to [Graf et al., 2007, Proposition 1], for $i \neq j$, we have $P\left(V_i(\mathbf{c}^*) \cap V_j(\mathbf{c}^*)\right) = 0$, so that with a slight abuse we may thought $R_P(\mathbf{c}^*)$ as the

distortion of the quantizers Q^* that maps each $V_j(\mathbf{c}^*)$ onto c_j^*, that is $R_P(Q^*) = P(du)\|u - Q^*(u)\|^2$.
As in the proof of Proposition 2.8, let $\Delta_{\infty,K} = \max_{x\in K} \mathrm{d}_{\mathbf{c}^*}(x)$. For every $j \in [\![1,k]\!]$, we let $\mathbf{c}^{*,-j}$ denote the $(k-1)$-points codebook obtained via removing c_j^* from $\mathbf{c}^*$. Let $i \in [\![1,k]\!]$ be fixed. Since, for all $j \in [\![1,k]\!]$, $V_j(\mathbf{c}^*) \cap K \neq \emptyset$, there exists x_1 and x_2 in K such that $(\mathrm{d}_{\mathbf{c}^{*,-i}} - \mathrm{d}_{c_i^*})(x_1) \leq 0$ and $(\mathrm{d}_{\mathbf{c}^{*,-i}} - \mathrm{d}_{c_i^*})(x_2) \geq 0$. Since K is connected, there exists $x_3 \in K$ such that $\mathrm{d}_{\mathbf{c}^{*,-i}}(x_3) = \mathrm{d}_{c_i^*}(x_3)$. It follows that, for every $x \in K \cap V_i(\mathbf{c}^*)$,

$$\begin{aligned}\mathrm{d}_{\mathbf{c}^{*,-i}}(x) &\leq \|x - x_3\| + \mathrm{d}_{\mathbf{c}^{*,-i}}(x_3)\\ &\leq \|x - c_i^*\| + 2\|x_3 - c_i^*\|\\ &\leq 3\Delta_{\infty,K}.\end{aligned}$$

Let $i_0 \in \arg\min_{i=1,\dots,k} P(V_i(\mathbf{c}^*))$. It holds $P(V_{i_0}(\mathbf{c}^*)) \leq k^{-1}$, so that

$$R_P(\mathbf{c}^{*,-i_0}) - R_P(\mathbf{c}^*) \leq \frac{9\Delta_{\infty,K}^2}{k}.$$

Next, let $x_0 \in K$ be such that $\mathrm{d}_{\mathbf{c}^*}(x_0) = \Delta_{\infty,K}$. With a slight abuse of notation we denote by $\mathbf{c}^{*,-i_0} \cup \{x_0\}$ the k-points codebook obtained via concatenation. Note that, for any $y \in \bar{\mathrm{B}}\left(x_0, \frac{\Delta_{\infty,K}}{4}\right)$, we have $\mathrm{d}_{\mathbf{c}^*}(y) \geq \frac{3\Delta_{\infty,K}}{4}$. This entails

$$\begin{aligned}R_P(\mathbf{c}^{*,-i_0} \cup \{x_0\}) &\leq R_P(\mathbf{c}^{*,-i_0}) - P\left(\bar{\mathrm{B}}\left(x_0, \frac{\Delta_{\infty,K}}{4}\right)\right)\left[\frac{9}{16}\Delta_{\infty,K}^2 - \frac{1}{16}\Delta_{\infty,K}^2\right]\\ &\leq R_P(\mathbf{c}^{*,-i_0}) - \frac{\Delta_{\infty,K}^2}{2}(a\Delta_{\infty,K}^b \wedge 1),\end{aligned}$$

whenever $\Delta_{\infty,K}^2 \leq 4r_0$. According to Proposition 2.8, $\Delta_{\infty,K} \to 0$ as $k \to +\infty$. This yields, for k large enough,

$$R_P(\mathbf{c}^{*,-i_0} \cup \{x_0\}) \leq R_P(\mathbf{c}^*) + \Delta_{\infty,K}^2\left(\frac{9}{k} - \frac{a\Delta_{\infty,K}^b}{2}\right).$$

Since $R_P(\mathbf{c}^{*,-i_0} \cup \{x_0\}) \geq R_P(\mathbf{c}^*)$, it holds

$$\Delta_{\infty,K} \leq C(ak)^{-\frac{1}{b}},$$

for k large enough. □

Theorem 2.9 assesses that optimal coverings of K with k points may be performed via quantization, in the noise-free case. Contrary to Proposition 2.8, this results does not easily adapts to Wasserstein perturbations, so that an open question is to know whether a $O(k^{-\frac{1}{b}} + \sigma)$ is attainable for a k-points optimal quantizer of Q. Note that the results of Theorem 2.9 hold for more general distance-based risk functions. The interested reader is referred to Graf and Luschgy [2000], Gruber [2004], Clarkson [2006] for an introduction to quantization in a more general framework.
Moving back to our initial problem, that is building a coreset for d_K based on a noisy n-sample $\mathbb{X}_n$, it remains to build optimal codebooks $\hat{\mathbf{c}}_n$ from sample, and to connect $\|\mathrm{d}_{\hat{\mathbf{c}}_n} - \mathrm{d}_K\|_\infty$ to $\|\mathrm{d}_{\hat{\mathbf{c}}_n} - \mathrm{d}_K\|_{\infty,K}$.
The first point may be carried out using empirical risk minimization, that is choosing

$$\hat{\mathbf{c}}_n \in \arg\min_{\mathbf{c}\in(\mathbb{R}^D)^k} Q_n(du)\mathrm{d}_{\mathbf{c}}^2(u),$$

where Q_n is the distribution of the sample $\mathbb{X}_n$, or other k-means like approach. More details are to be found in the following Chapter 3, let us briefly sum up here the convergence rates for $R_Q(\hat{\mathbf{c}}_n) - R_Q(\mathbf{c}_Q^*)$, depending on assumptions on Q.

Q	SubGaussian	Bounded support	Margin condition
Rate	$(\log(n)^3/n)^{\frac{1}{2}}$	$n^{-1/2}$	n^{-1}
Upper	Brécheteau and Levrard [2020]	Biau et al. [2008], Linder [2002]	Levrard [2015, 2018]
Lower	\	Bartlett et al. [1998], Antos [2005], Levrard [2015]	Levrard [2018]

Table 2.1: Convergence rates for $\mathbb{E}(R_Q(\hat{\mathbf{c}}_n) - R_Q(\mathbf{c}^*_Q))$.

Some of the results exposed in Table 2.1 may be retrieved from results in Bachem et al. [2017]. Note also that other conditions on Q might ensure a n^{-1} rate, such as the one exposed in Pollard [1982] or Antos et al. [2005]. Some insights on the connections between these conditions and the aforementioned ones may be found in Levrard [2013, 2015, 2018].
Connecting $R_Q(\hat{\mathbf{c}}_n)$ with $R_P(\hat{\mathbf{c}}_n)$ may be performed using (2.1). Then, using Proposition 2.8, Item (ii) allows to bound $\|\mathrm{d}_{\hat{\mathbf{c}}_n} - \mathrm{d}_K\|_{\infty,K}$ whenever P is (a,b)-standard.
The second point (connecting $\|\mathrm{d}_{\hat{\mathbf{c}}_n} - \mathrm{d}_K\|_\infty$ to $\|\mathrm{d}_{\hat{\mathbf{c}}_n} - \mathrm{d}_K\|_{\infty,K}$) is hopeless in the general case, as alleged by the following example.

Example 2.10 *: [Brécheteau and Levrard, 2020, Example 3].* Let $Q_\beta = \beta \mathcal{U}_{S(0,1)} + (1-\beta)\mathcal{U}_{B(0,1)}$ be a noisy version of $P = \mathcal{U}_{S(0,1)}$, the uniform distribution on the circle, for some $\beta \in (0,1)$.
Then $W_2(Q_\beta, P) \leq \sqrt{1-\beta}$, so that, if $\mathbf{c}^*$ denotes an optimal k-points codebook for Q_β, we have, for k large enough,

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

Nonetheless, for every $\rho > 0$, there exists $k_{\rho,\beta}$ such that, for all $k \geq k_{\rho,\beta}$, $\mathbf{c}^*$ has at least one codepoint in $B(0,\rho)$.
As a consequence,

$$\sup_{k \geq 0} \|\mathrm{d}_{S(0,1)} - \mathrm{d}_{\mathbf{c}^*}\|_\infty \geq \sup_{k \geq 0} |\mathrm{d}_{S(0,1)}(0) - \mathrm{d}_{\mathbf{c}^*}(0)| = 1.$$

The intuition behind Example 2.10 is that though optimal codebooks designed via classical quantization can yield provably good covering of topological structures such as manifolds, they are also likely to have codepoints far from the structure in some noisy cases. In this case, geometric inference based on the sublevel sets of $\mathrm{d}_{\mathbf{c}^*}$ might lead to poor results. From a technical point of view, the good covering properties of optimal codebooks exposed in Proposition 2.8 only provide bounds on

$$\sup_{x \in \mathbb{R}^D} \mathrm{d}_{\mathbf{c}^*}(x) - \mathrm{d}_K(x).$$

Example 2.10 illustrates that $\sup_{x \in \mathbb{R}^D} \mathrm{d}_K(x) - \mathrm{d}_{\mathbf{c}^*}(x)$ may stay arbitrarily large whenever $k \to +\infty$.
Note however that for special regimes of k and σ, such as $\sigma << k^{-1} << \tau_K$, the averaging effect of optimal codebooks might still lead to suitable approximations of d_K. Such approaches could lead to case-dependent bounds on $\|\mathrm{d}_K - \mathrm{d}_{\mathbf{c}^*}\|_\infty$, that will not be discussed here. To bypass the spurious codepoints issue, we rather adopt a robustification method, as detailed in the following section.

2.3 Robust quantization via k-PDTM

To prevent optimal codebooks obtained via classical quantization of the noisy distribution Q to have code points far from the targeted support K, a first idea could be to remove code points whose Voronoi cell has too small weight (thus accounting for noise). The intuition behind is that, to account for noise, one may not target the support of the distribution (area with positive mass) but rather areas of $\mathbb{R}^D$ that gather enough mass. In other words, we want to estimate *surrogates* of the support that we hope close enough to the original support.
Such surrogates estimation approaches has been successfully developed in the manifold case. In Genovese et al. [2014], for the noise model $Q \sim P \star \mathcal{N}(0, \sigma^2)$, the targeted surrogate is the so-called *ridge* R_σ of Q, that is the set of d-maxima of the density f_Q. In the $\mathcal{P}^{2,d,D}_{\tau_{\min},\infty}(f_{\min}, f_{\max})$ model, [Genovese et al., 2014, Theorem 7] ensures that $\mathrm{d_H}(R_\sigma, M) = O\left(\sigma^2 \log(\sigma^{-1})\right)$, so that their ridge estimation procedure provides an estimator $\hat{R}$ such that

$$\mathrm{d_H}(\hat{R}, M) \lesssim \sigma^2 \log(\sigma^{-1}) + \left(\frac{\log(n)}{n}\right)^{\frac{2}{D+8}},$$

that is enough to allow for topological inference whenever $\sigma^2 << \tau_M$. Note that the proposed estimator rely on differential quantities of density estimator, this requires at least estimation of tangent spaces. Such an estimator might be "sparsified", for instance using the Farthest Point Sampling algorithm onto the input sample $\mathbb{X}_n$, allowing to reach a prescribed accuracy level with less points.
The tangential structure is also heavily used in Fefferman et al. [2018], where the authors choose as surrogate a somehow local linear interpolation of Q, showing clear connection with the work in Maggioni et al. [2016], Canas et al. [2012] in the additive noise model. This locally linear approximation can be proved $O(\sigma d^7 \sqrt{D})$ close to M, and can be build with $O((\sigma\sqrt{D})^{-d})$ points. Thus, the proposed (inconsistent) estimator conveys the correct topology whenever $\sigma << \tau_M$ and $n \gtrsim \sigma^{-d}$, with the additional benefit of a much milder dependency on D than the method proposed in Genovese et al. [2014].
It is worth noting that the two aforementioned methods are valid under the manifold assumption in the additive noise case, and both rely on tangent spaces estimation. In Brécheteau and Levrard [2020], we propose a more general approach that is valid under the (a, b)-standard model and Wasserstein perturbation.

2.3.1 The distance to measure (DTM)

Our approach is based on a particular surrogate of d_K that originates in Chazal et al. [2011], the Distance to Measure (DTM). For a distribution P such that $P(dx)\|x\|^2 < \infty$, and a mass parameter $h > 0$, we denote by $\mathcal{P}_h(P)$ the set of h-submeasures of P, that is

$$\mathcal{P}_h(P) = \left\{ \frac{1}{h} Q \mid Q << P \quad \text{and} \quad Q(\mathbb{R}^D) = h \right\}. \tag{2.2}$$

The distance to measure may be defined as follows.

Definition 2.11 :

> Let P be such that $P(dx)\|x\|^2 < \infty$. Then, for $x \in \mathbb{R}^D$, the *distance to the measure* P at x is defined by
>
> $$\mathrm{d}_{P,h}(x) = \inf_{Q \in \mathcal{P}_h(P)} Q(du)\, \|u - x\|^2 .$$

Equivalent definitions of the DTM may be stated, that reveal its stability with respect to Wasserstein perturbations. Namely, if $r_{P,h}(x) = \inf\{r \geq 0 \mid P\left(\bar{\mathrm{B}}(x,r)\right) \geq h$, it holds

$$\mathrm{d}^2_{P,h}(x) = \frac{1}{h}\int_0^h r^2_{P,u}(x)du,$$

so that, for a point x with $\mathrm{d}_K(x) > 0$ (where $K = \mathrm{Supp}(P)$), if $Q = \beta P + (1-\beta)N$ for $(1-\beta) < h$, we have $\mathrm{d}_{Q,h}(x) \geq \mathrm{d}_K(x)$ (see Figure 2.4). From a formal point of view, the

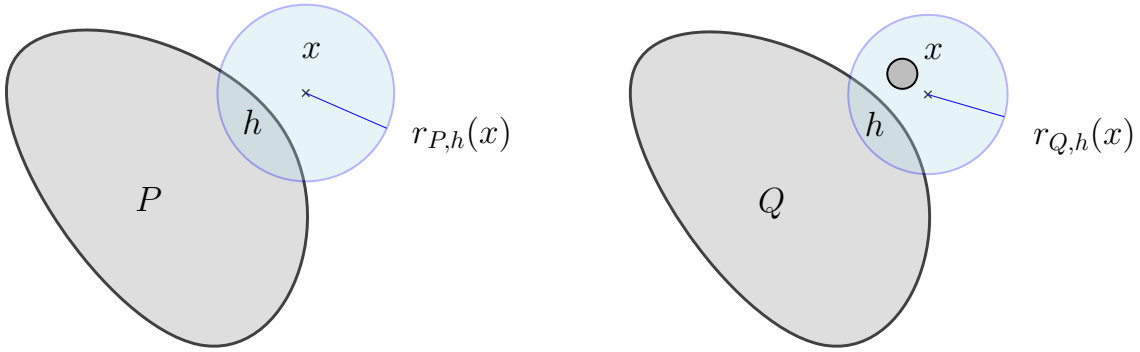


Figure 2.4: To grab enough mass from P, $r_{Q,h} \geq \mathrm{d}_K(x)$ whenever $(1-\beta) < h$.

desired Wasserstein stability derives from the following characterization.

THEOREM 2.12 : [CHAZAL ET AL., 2011, PROPOSITION 3.3 AND THEOREM 3.5]

Let P be such that $P(du)\|u\|^2 < \infty$, and $h > 0$. Then

$$\mathrm{d}^2_{P,h}(x) = \inf_{Q \in \mathcal{P}_h(P)} W_2^2(Q, \delta_{\{x\}}).$$

As a consequence, if $Q(du)\|u\|^2 < \infty$, we have

$$\|\mathrm{d}_{Q,h} - \mathrm{d}_{P,h}\|_\infty \leq \frac{W_2(P,Q)}{\sqrt{h}}.$$

Now, for the DTM to consist in a good surrogate of d_K, it remains to connect $\mathrm{d}_{P,h}$ to d_K, whenever $P = \mathrm{Supp}(K)$. It is immediate that $\mathrm{d}_{P,h} \geq \mathrm{d}_K$, in fact d_K may be thought of as the limiting case $\lim_{h\to 0} \mathrm{d}_{P,h}$. A converse bound may be stated in the case of (a,b)-standard measures.

PROPOSITION 2.13 : [CHAZAL ET AL., 2011, FROM LEMMA 4.7]

Assume that P is (a,b)-standard at scale r_0, and denote by $K = \mathrm{Supp}(P)$. Then, for any $h \leq a r_0^b$, we have

$$\|\mathrm{d}_{P,h} - \mathrm{d}_K\|_\infty \leq \left(\frac{h}{a}\right)^{\frac{1}{b}}.$$

Moreover, for every $h \in (0,1)$, it holds

$$\|\mathrm{d}_{P,h} - \mathrm{d}_K\|_\infty \leq \frac{\mathrm{Diam}(K)}{r_0}\left(\frac{h}{a}\right)^{\frac{1}{b}}.$$

Thus, combining Theorem 2.12 and Proposition 2.13 ensures that the topology of K may be retrieved from the sublevel sets of $\mathrm{d}_{Q,h}$ whenever $h \sim \tau_K^b$ and $W_2(P,Q) \lesssim \tau_K^{1+\frac{b}{2}}$, in the (a,b)-standard case. The Distance to Measure then appears as a suitable surrogate for the distance to support for the (a,b)-standard case with Wasserstein perturbation. As noted by Guibas et al. [2013], in practice, if Q_n denotes the empirical distribution induced by $\mathbb{X}_n$, then computing the sublevel sets of the empirical DTM would require to build unions of $\binom{n}{q}$ balls, where $q = hn$. Computing the corresponding topological features becomes intractable for large n's, hence the particular interest of k-points approximations of the DTM. Note however that Guibas et al. [2013], Buchet et al. [2016] propose approximations of the DTM, whose sublevel sets consists of union of n balls for the empirical DTM, rendering the computation of topological features tractable for moderate sample sizes. In what follows, we intend to build k-points coreset for the DTM, that would lead to compute the homology of a union of k balls.

2.3.2 A coreset for the DTM

To build a k-points approximation of the DTM, first note that it may be expressed as a power distance.

Lemma 2.14 : [Brécheteau and Levrard, 2020, Lemma 11]

Let P be such that $P(du)\|u\|^2 < \infty$, $h > 0$ and $x \in \mathbb{R}^D$. Then

$$\mathrm{d}^2_{P,h}(x) = \min_{Q \in \mathcal{P}_h(P)} \left[\|x - m(Q)\|^2 + v(Q) \right] = \min_{\tau \in \mathbb{R}^D} \left[\|x - \tau\|^2 + \omega^2_{P,h}(\tau) \right],$$

where $m(Q)$ and $v(Q)$ respectively denote the mean and variance of Q, and

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

In particular, the elements of $\mathcal{P}_h(P)$ that achieves the minimum are the elements of

$$\mathcal{P}_{x,h}(P) = \left\{ \frac{1}{h} Q \mid Q_{|\mathrm{B}(x, r_{p,h}(x))} = P_{|\mathrm{B}(x, r_{p,h}(x))} \quad \text{and} \quad \mathrm{Supp}(Q) \subset \bar{\mathrm{B}}(x, r_{P,h}(x)) \right\}.$$

Lemma 2.14 ensures that $\mathrm{d}^2_{P,h}(x)$ may be expressed as $\|x - c\|^2 + \omega^2_{P,h}(c)$ or $\|x - m(P_{t,h})\|^2 + v(P_{t,h})$, for a suitable c or t, depending on the parametrization. In particular, Lemma 2.14 ensures that the DTM is a distance-like function in the sense of Definition 2.1. Approximating $\mathrm{d}^2_{P,h}$ via a k-points based distance may then be processed in a straightforward way, by considering

$$\min_{j=1,\dots,k} \|x - c_j\|^2 + \omega^2_{P,h}(c_j)$$

or

$$\min_{j=1,\dots,k} \|x - m(P_{t_j,h})\|^2 + v(P_{t_j,h}),$$

for a suitable choice of $\mathbf{c} = (c_1, \dots, c_k)$ or $\mathbf{t} = (t_1, \dots, t_k)$. As explained at the beginning of Section 2.2, choosing $\mathbf{c}$ via a regular grid of K, though theoretically optimal, does not adapt to the noisy case. We rather adopt a quantization approach, that is we choose $\mathbf{t}$ (or $\mathbf{c}$) via the minimization of a L_2 criterion.

Definition 2.15 :

Let P be such that $P(du)\|u\|^2 < \infty$, and $h > 0$. For all $x \in \mathbb{R}^D$, the *k-PTDM* at x is defined by

$$\mathrm{d}^2_{P,h,k}(x) = \min_{j=1,\dots,k} \|x - c_j^*\|^2 + \omega^2_{P,h}(c_j^*) = \min_{j=1,\dots,k} \|x - m(P_{t_j^*,h})\|^2 + v(P_{t_j^*,h}),$$

where

$$\mathbf{c}^* \in \arg\min_{\mathbf{c}\in(\mathbb{R}^D)^k} P(du) \min_{j=1,\dots,k} \|u - c_j\|^2 + \omega^2_{P,h}(c_j),$$

or equivalently

$$\mathbf{t}^* \in \arg\min_{\mathbf{t}\in(\mathbb{R}^D)^k} P(du) \min_{j=1,\dots,k} \|u - m(P_{t_j,h})\|^2 + v(P_{t_j,h}).$$

A proof of the equivalence between the two parametrizations is given by [Brécheteau and Levrard, 2020, Theorem 12]. Intuitively, elements c such that $\omega_{P,h}(c)$ is small will be favored. Such c's gather a proportion h of the mass of P on their neighborhood. On the contrary, for elements c such that $\omega_{P,h}(c)$ is large, the ball associated to such c's will appear in the r-sublevel set of the function $x \mapsto \|\cdot - c\|^2 + \omega^2_{P,h}(c)$ for large r's only. This is a way to "remove" code points that fall far from the support.
According to Definition 2.15, the squared k-PDTM may be interpreted as the closest squared k-power distance to the squared DTM from above, in terms of $L_1(P)$ norm. This interpretation comes from the straightforward inequality $\|x - c\|^2 + \omega^2_{P,h}(c) \geq \mathrm{d}^2_{P,h}(x)$. The resulting inequality $\mathrm{d}^2_{P,h,k} \geq \mathrm{d}^2_{P,h}$ allows for further comparison with k-means approximation of the distance-to-compact-set in noisy settings. A follow-up of Example 2.10 is exposed below.

Example 2.16 *: Example 2.10, follow-up.* For the distribution $Q_\beta = \beta\mathcal{U}_{S(0,1)} + (1-\beta)\mathcal{U}_{B(0,1)}$. If $h > 1 - \beta$, since $\mathrm{d}^2_{Q_\beta,h,k}(0) \geq \mathrm{d}^2_{Q_\beta,h}(0)$, we have $\mathrm{d}^2_{Q_\beta,h,k}(0) \geq 1 - \frac{1-\beta}{h}$. As a consequence, $\inf_{k\geq 0} \mathrm{d}^2_{Q_\beta,h,k}(0) \geq 1 - \frac{1-\beta}{h}$, whereas $\inf_{k\geq 0} \mathrm{d}^2_{\mathbf{c}^*}(0) = 0$, where $\mathbf{c}^*$ denotes an optimal k-points codebook.

The above example shows that we can expect the k-PDTM to approximate well the distance-to-compact-set in remote areas, contrary to the distance based on classical quantization, $\mathrm{d}_{\mathbf{c}^*}$. To formally assess the relevance of our approach, it remains to establish approximation properties of the k-PDTM in the (a, b)-standard case, as well as its robustness with respect to Wasserstein perturbations.
Robustness of the k-PDTM is assessed by the following result.

Proposition 2.17 : [Brécheteau and Levrard, 2020, Proposition 17]

Let P be such that $\mathrm{Supp}(P) \subset \bar{\mathrm{B}}(0, R)$, for some $R > 0$, and Q be such that $Q(du)\|u\|^2 < \infty$. Let $\mathrm{d}^2_{Q,h,k}$ denote a k-PDTM for Q. Then, $P(du)\left|\mathrm{d}^2_{Q,h,k} - \mathrm{d}^2_{P,h}\right|(u) \leq B_{P,Q,h,k}$, where

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

The first and third term in the above equation might be bounded in term of Wasserstein

distance. Indeed, using [Chazal et al., 2011, Theorem 3.5], we easily have

$$\|\mathrm{d}^2_{Q,h} - \mathrm{d}^2_{P,h}\|_{\infty,\mathrm{B}(0,R)} \leq \frac{\sigma^2}{h} + \frac{4R\sigma}{\sqrt{h}},$$

where $W_2(P,Q) \leq \sigma$. Next, the third term might be bounded by $4\sigma R$. The second term is less usual, and depends on the approximation properties of the k-PDTM. Such an approximation term might be roughly bounded in the (a,b)-standard case.

Proposition 2.18 :

Assume that P is (a,b)-standard at scale r_0. Then, for $k \geq \frac{r_0^{-b}}{a}$,

$$P(du)\mathrm{d}^2_{P,h,k}(u) \leq \frac{4R^2}{r_0^2}\left(\frac{h}{a}\right)^{\frac{2}{b}} + 4(ak)^{-\frac{2}{b}}.$$

Proposition 2.18 easily follows from the proof of Proposition 2.8, Item (i) and Proposition 2.13. With more structural assumptions, refined bounds on $P(du)(\mathrm{d}^2_{P,h,k} - \mathrm{d}^2_{P,h})(u)$ may be derived.

Proposition 2.19 : [Brécheteau and Levrard, 2020, Corollary 16]

Let $P \in \mathcal{C}^{2,d,D}_{\tau_{\min},\infty}(f_{\min}, f_{\max})$, and suppose that $\mathrm{Supp}(P) \subset \bar{\mathrm{B}}(0,R)$. Then, for $k \geq c_{\tau_{\min},f_{\min}}$ and $h \leq C_{\tau_{\min},f_{\min}}$, we have

$$0 \leq P\mathrm{d}^2_{P,h,k} - \mathrm{d}^2_{P,h} \leq C_{\tau_{\min},f_{\min},f_{\max}} k^{-2/d}.$$

Proposition 2.19 is a particular instance of a more general result [Brécheteau and Levrard, 2020, Proposition 14] that relates the L_2 bias of the k-PDTM to covering numbers and stability of local means. Note that Proposition 2.19 allows to bound $\|\mathrm{d}_{P,h,k} - \mathrm{d}_{P,h}\|_\infty$ in a similar way to that of Proposition 2.8. This proves that the k-PTDM is a coreset for the DTM in this case. However, the simpler bound given by Proposition 2.18 is enough to provide guarantees on $\|\mathrm{d}_K - \mathrm{d}_{P,h,k}\|_\infty$ in the (a,b)-standard case.

Proposition 2.20 : [Brécheteau and Levrard, 2020, Proposition 18]

Let P be (a,b)-standard at scale r_0, with $K = \mathrm{Supp}(P) \subset \bar{B}(0,R)$, and let Q be such that $W_2(P,Q) \leq \sigma$. Then, we have

$$\|\mathrm{d}_{Q,h,k} - \mathrm{d}_K\|_\infty \leq \max\left\{2\left(\frac{ar_0^b}{(2R)^b}\right)^{-\frac{1}{b+2}} \Delta_P^{\frac{2}{b+2}},\ 2\Delta_P,\ \sigma h^{-\frac{1}{2}}\right\},$$

where $\Delta_P^2 = P(du)\mathrm{d}^2_{Q,h,k}(u)$. Besides, if $\sigma \leq R$, it holds, for $k \geq \frac{r_0^{-b}}{a}$,

$$\Delta_P \leq C\left(\frac{R}{r_0}\left(\frac{h}{a}\right)^{\frac{1}{b}} + (ak)^{-\frac{1}{b}} + \sqrt{\frac{R\sigma}{h}}\right).$$

It is worth noting that the ambient dimension D plays no role in Proposition 2.20 To fix ideas, let us consider the (a,b)-standard case with $b \geq 1$, $R, \tau_K \leq 1$. In this case, for the

sublevel sets of $\mathrm{d}_{Q,h,k}$ to provide the correct topology, one has to choose

$$h \sim \tau_K^{\frac{b(b+2)}{2}}, \quad \sigma \lesssim \tau_K^{\frac{3b(b+2)}{2}}, \quad k \gtrsim \tau_K^{-\frac{b(b+2)}{2}}.$$

These conditions are more stringent than the regime $h \sim \tau_K^b$ and $\sigma \lesssim \tau_K^{1+\frac{b}{2}}$ that is sufficient for the DTM to provide topological guarantees. As well, the $\tau_K^{-\frac{b(b+2)}{2}}$ bound for k is in line with Proposition 2.8, though worse than the $\tau_K^{-\frac{1}{b}}$ that follows from Theorem 2.9 in the noiseless case. It is an open question to know whether the optimal relation between $\|\mathrm{d}_{\mathbf{c}^*} - \mathrm{d}_K\|_{L_2(P)}$ and $\|\mathrm{d}_{\mathbf{c}^*} - \mathrm{d}_K\|_\infty$ exposed in Theorem 2.9 could be adapted to the k-PDTM.
Once the k-PTDM is proved to be a coreset for d_K under (a,b)-standard assumptions, it remains to provide an estimate of the k-PDTM from sample.

2.3.3 Sample-based approximation

Recall that $\mathbb{X}_n$ is an n-sample from Q, with empirical distribution $Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. The empirical k-PDTM is just the straightforward plug-in $\mathrm{d}_{Q_n,h,k}$, that intend to approximate $\mathrm{d}_{Q,h,k}$. Note that for $k = n$, the empirical n-PDTM is the q-witnessed distance of Guibas et al. [2013], for $q = nh$. The following result provides bounds on the estimation error between $\mathrm{d}_{Q_n,h,k}$ and $\mathrm{d}_{Q,h,k}$, in a sub-Gaussian additive noise model. We formalize our definition of sub-Gaussian random variable below.

Definition 2.21 :

> Let X be a random variable on $\mathbb{R}^D$. X is called *sub-Gaussian with variance* σ^2 if, for every $t > 0$,
>
> $$\mathbb{P}\left(\|X - \mathbb{E}(X)\| \geq t\right) \leq 2e^{-\frac{t^2}{2\sigma^2}}.$$

Theorem 2.22 : [Brécheteau and Levrard, 2020, Theorem 19]

> *Let P be such that* $\mathrm{Supp}(P) \subset \bar{\mathrm{B}}(0,R)$. *Assume that, for $i \in [\![1,n]\!]$, $X_i = Y_i + Z_i$, where the (Y_i)'s and (Z_i)'s are independent, the Y_i's are sampled from P, and the Z_i's from a sub-Gaussian distribution with variance σ^2, with $\sigma \leq R$. Then, for any $p > 0$, with probability larger than $1 - 10n^{-p}$, we have*
>
> $$\left|P(du)(\mathrm{d}^2_{Q_n,h,k} - \mathrm{d}^2_{Q,h,k})(u)\right| \leq C\sqrt{kd}\frac{R^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C\frac{R\sigma}{\sqrt{h}}.$$

This $\sqrt{kd/n}$ rate is in line with the rate given in Linder [2002] in a classical quantization with bounded support framework. Note that the second term of the bounds is due to the integration with respect to P, the first term of the right-hand side being a bound on $\left|Q(du)(\mathrm{d}^2_{Q_n,h,k} - \mathrm{d}^2_{Q,h,k})(u)\right|$. It is also worth mentioning that the result of Theorem 2.22 is valid under the milder assumptions $X_i \sim Q$, Q is sub-Gaussian with variance bounded by $(R+\sigma)^2$, and $W_2(P,Q) \leq \sigma \leq R$. Combining the results of Theorem 2.22 and Proposition 2.18 ensures that, whenever $n \gtrsim k^{\frac{b+4}{b}}$, and h, σ, k are properly chosen (see below Proposition 2.18), topological guarantees hold on the sublevel sets of $\mathrm{d}_{Q_n,h,k}$, with high probability. Conversely, choosing $k \sim n^{\frac{b}{b+4}}$ yields an optimal bias-variance tradeoff in the aforementioned bounds.

Whether n^{-1} rates might be achieved as for the classical empirical quantization scheme via k-means, under regularity assumptions such as those depicted in Levrard [2015], is still an open question. However, optimality of the $1/\sqrt{n}$ rate over the class of bounded distributions may be proved similarly to Bartlett et al. [1998].

PROPOSITION 2.23 : [BRÉCHETEAU AND LEVRARD, 2020, PROPOSITION 21]

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

$$\mathrm{d}^2_{P,h,\mathbf{t}} : x \mapsto \min_{j \in [\![1,k]\!]} \left[\|x - m(P_{t_j,h})\|^2 + v(P_{t_j,h}) \right].$$

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

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

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

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

From a practical point of view, minimizing $\mathbf{c} \mapsto Q_n(du) \min_{j=1,\dots,k} \|u - c_j\|^2 + \omega^2_{Q_n,h}(c_j)$ seems intractable (for $h = 0$, that is the k-means case, the problem is provably NP-hard for $D \geq 2$ Mahajan et al. [2012]). In the k-means case, local minimizers of the objective function might be found via a iterative schemes, such as Lloyd's algorithm Lloyd [1982] or Mac-Queen algorithm MacQueen [1967]. These algorithms have been extended to the case where the squared Euclidean norm is replaced by a Bregman divergence in Banerjee et al. [2005b]. In a nutshell, Bregman divergences are functions of the form $\mathrm{d}_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla_y \phi, x - y \rangle$, where ϕ is a convex function. The formal definition of Bregman divergences will be given in Section 3.1. To apply the iterative scheme of Banerjee et al. [2005b] to our case, it remains to express $c \mapsto \min_{j=1,\dots,k} \|u - c_j\|^2 + \omega^2_{Q_n,h}(c_j)$ as a Bregman divergence. To do so, denote by

$$\psi_{P,h} : \begin{cases} \mathbb{R}^D & \to & \mathbb{R} \\ x & \mapsto & \|x\|^2 - \mathrm{d}^2_{P,h}(x). \end{cases}$$

According to [Chazal et al., 2011, Proposition 3.6], $\psi_{P,h}$ is convex, with subgradients $\{m(P_{x,h}) \mid P_{x,h} \in \mathcal{P}_{x,h}(P))\}$. Hence, straightforward computation leads to

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

for any $x, t \in \mathbb{R}^D$. Thus, minimizing $\mathbf{c} \mapsto Q_n(du) \min_{j=1,\dots,k} \|u - c_j\|^2 + \omega^2_{Q_n,h}(c_j)$ is equivalent to minimize $\mathbf{c} \mapsto Q_n(du) \min_{j=1,\dots,k} \mathrm{d}_{\psi_{Q_n,h}}(u, c_j)$, that falls in the scope of [Banerjee et al., 2005b, Proposition 2]. More details on the algorithm are given in Section 3.1.

2.3.4 Numerical illustration

According to Theorem 2.2, if K is a compact subset of $\mathbb{R}^D$, then geometric and topological information about K can be recovered from some r-sublevel sets of d_K.

To assess the relevancy of our approach in a noisy topological inference setting, we will compute the persistence diagrams associated with the empirical k-PDTM, and compare it with the outputs of other methods.
Following Guibas et al. [2013], we choose for K the infinity symbol embedded in $\mathbb{R}^2$. The targeted persistence diagram associated to d_K is the one depicted in Figure 2.1 (right). This diagram contains one red point $(0, \infty)$, that corresponds to the connected component (0-dimensional topological feature), and two green points that account for the two holes (1-dimensional topological features). We generated a sample of 200 points, uniformly on the infinity symbol, with an additional additive Gaussian noise, with standard deviation $\sigma = 0.02$. This sample is corrupted by 80 outliers – 40 points generated according to the uniform distribution on the rectangle $[-2, 5] \times [-2, 2]$ and 40 points on the rectangle $[-4, 7] \times [-4, 4]$. This results in a corrupted sample $\mathbb{X}_n$ of 280 points.
We compared three methods to recover relevant features of M from $\mathbb{X}_n$. Each method boils down to build a coreset approximation f of d_M in the sense of Definition 2.6: they are of the type $f : x \mapsto \sqrt{\min_{i \in I} \|x - \tau_i\|^2 + \omega_i^2}$, for some finite set I, centers $\tau_i \in \mathbb{R}^d$ and weights $\omega_i \geq 0$. The first function we consider is derived from the k-means algorithm McLachlan and Peel [2000] ($|I| = k$, centers τ_i are given by the optima of the k-means criterion and $\omega_i = 0$), the second is the q-witnessed distance Guibas et al. [2013] ($|I| = n = 280$, it coincides with the k-PDTM for $k = n$, with mass parameter $h = q/n$) and the third one is the k-PDTM ($|I| = k$, with mass parameter $h = q/n$).
These methods depend on two parameters q and k, that are respectively fixed to $q = 10$ and $k = 50$. Roughly, q is chosen small enough so that the distance to the q-th nearest neighbor remains small compared to the curvature of M but large enough to deal with noise, and k is chosen large enough so that a uniform grid with k points has grid size small compared to the curvature of M. More details on this heuristic can be found in the Appendix of Brécheteau and Levrard [2020].

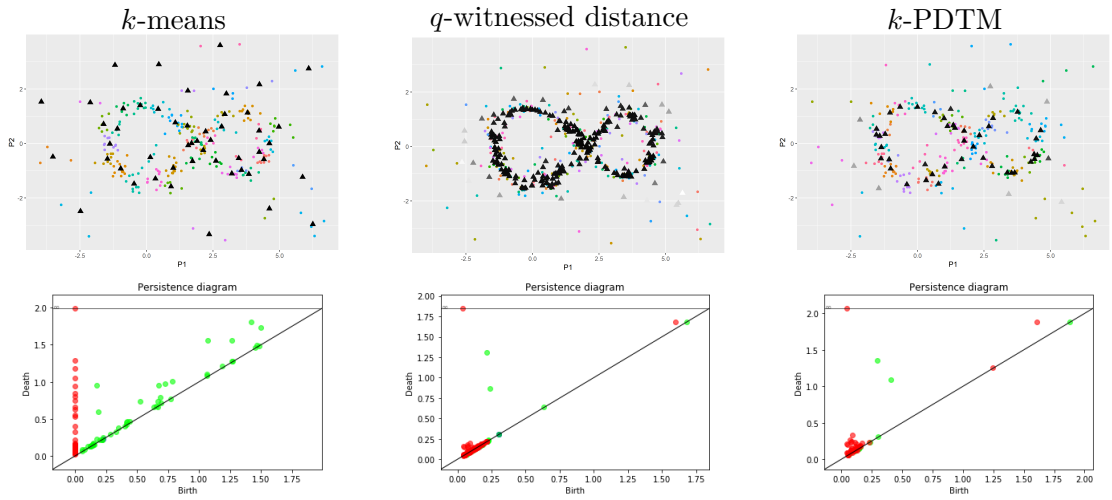


Figure 2.5: Centers and persistence diagrams for the noisy infinite symbol and three algorithms.

We implemented the three methods with the `R` software. We used the `R` function `kmeans` for the first method, and the `FNN R` library to compute the nearest neighbors for the two other methods. In Figure 2.5 we plotted the points of $\mathbb{X}_n$. Points are represented with the same color when they lie in the same weighted Voronoi cell (for the centers τ_i and weights ω_i^2). Centers τ_i are represented by triangles and colored in function of the weights ω_i^2 (black centers correspond to $\omega_i^2 = 0$). The second row of Figure 2.5 depicts the corresponding persistence diagrams, for the three methods. They were obtained

using the function `weighted_alpha_complex_3d_persistence` in the `Gudhi C++` library, based on alpha-shapes Edelsbrunner [1992]. We observe on Figure 2.5 that the three main features of the symbol infinity (one connected component, two holes) are recovered for the k-PDTM and the q-witnessed distance, but not for k-means. As exposed in Example 2.10, this is due to the "void-filling" drawback of k-means.
It is worth noting that outliers points are somehow identified by our method: they correspond to high levels of $\hat{\mathrm{d}}$. As detailed in the following Section 3.1, it is possible to combine topological inference and outlier detection based on the k-PDTM.

2.4 Discussion and directions for future research

Towards structure-adaptive robust quantization.
Though optimal in the (a, b)-standard case (see Theorem 2.9), centroid-based quantization does not allow to take advantage of the structure of the support, for instance in the manifold case. In the noiseless case, the k-flats approach, that is replacing $\mathrm{d}_{\mathbf{c}}$ by the distance to $\bigcup_{j=1}^{k} \mathrm{B}_{T_j}(c_j, \lambda)$, for some subspaces T_j, yields improved bias term in the $\mathcal{C}^2$ manifold case (see, e.g., Canas et al. [2012], Maggioni et al. [2016]), as in the standard manifold estimation scheme. Pushing further the comparison, approximating d_M via the distance to a k-union of polynomial patches that minimize a L^2-like risk also improves the bias term. As for the manifold estimation case, investigating the dependency of the estimation error with respect to the regularity k may also lead to regularity-adaptive polynomial quantization schemes. Nevertheless, to convert L_2-type optimality results into optimal covering properties, structural results ensuring that centers of patches are well-separated (as in [Clarkson, 2006, Theorem 5.4]) should be needed, that are not established even in the k-flats case, up to our knowledge.
Robustification of the aforementioned structure adaptive methods is likely to need adaptation of the DTM principle. A natural extension can be based on optimal transportation: since $\mathrm{d}^2_{P,h}(u) = \inf_{x \in \mathbb{R}^D} W_2^2(P_{x,h}, \delta_{\{u\}})$, where $P_{x,h}$ is roughly $P_{|\mathrm{B}(x,r_h)}$, one could attempt to replace $P_{|\mathrm{B}(x,r_h)}$ by $P_{|\mathrm{B}(C(x,\lambda),r_h)}$, where $C(x, \lambda)$ denotes a polynomial cap of radius λ centered at x and $r_h = \inf\{r \geq 0 \mid P(\mathrm{B}(C(x, \lambda), r)) \geq h\}$. The quantization scheme would then be to minimize a criterion that looks like $P(du) \min_{j=1,\dots,k} P_{|\mathrm{B}(C(c_j,\lambda),r_j(h))}(dv)\|u - v\|^2$. Though purely theoretical, such an approach would benefit from robustness in terms of Wasserstein perturbations. Implementing it in practice would require to carry out optimization on the polynomial cap parameters, as well as the choice of a cap radius λ.
An other approach to extend the k-PDTM to k-flats case might be derived from a local maximum likelihood point of view. Namely, for a cell C_j associated with the measure $P_{t_j,h}$, the risk on C_j writes as

$$P_{t_j,h}(dv) P_{|C_j}(du) \left[-\log(V(u \mid e^{-\|\cdot - v\|^2})) \right] = P_{|C_j}(du) \left[-\log(V(u \mid e^{-\|\cdot - m_j\|^2})) + v_j \right],$$

where $V(. \mid f)$ denotes the likelihood with respect to f. This suggests that quantization with k-PTDM may be thought of as local matching with Gaussian spherical distributions in terms of Kullback distance. Thus, a natural extension that could take advantage of a structural anisotropy of P could be to minimize

$$P(du) \min_j P_{\mu_j,\Sigma_j,h}(dv) \left[-\log \left(V \left(u \mid e^{-\|\cdot - v\|^2_{\Sigma_j^{-1}}} \right) \right) \right],$$

where Σ_j ranges in a structured space of covariance matrices (for instance with $D - d$ eigenvalues fixed at σ^2, smaller than the d largest ones), and $P_{\mu_j,\Sigma_j,h}$ is a submeasure of P constrained on the corresponding Mahalanobis ball. This heuristic amounts to local

matching with heteroscedastic Gaussian distributions, that show clear connections with robust PCA. This is ongoing work with C. Brécheteau and B. Michel.
Let us conclude by stressing on the fact that whatever the chosen approach, a special attention has to be paid to complexity of the quantization algorithms, not to deprecate the computational benefits of the quantization-based inference.

Chapter 3

Outlier detection and Clustering

The last chapter presents two extensions and applications of the quantization principle, corresponding to the papers Brécheteau et al. [2021] and Chazal et al. [2021], Royer et al. [2021].
The first extension, exposed in Section 3.1, considers the general quantization with Bregman divergence framework, that intends to minimize $\mathbf{c} \mapsto P(du)\mathrm{d}_\phi(u, \mathbf{c})$, in the case where the sample may contain a prescribed proportion of outliers. After introducing some base models for outliers (Huber noise model and adversarial noise), as well as a way to seize robustness to noise in terms of Finite Sample Breakdown Point, we introduce our robust quantization scheme, based on the minimization of $\mathbf{c} \mapsto \min_{A|P(A)\geq 1-\alpha} P(du)\mathrm{d}_\phi(u, \mathbf{c})\mathbb{1}_A(u)$, that is a trimmed distortion minimization approach. Some theoretical guarantees on the distortion and robustness of empirical risk minimizers are given, that emphasize the influence of a so-called discernability factor, which seizes the minimal portion of a signal quantized with k points that can be considered as noise when quantized with $k-1$ points. We also provide a Lloyd-type algorithm, and investigates the numerical properties of our robust quantization scheme in two case studies: robust topological inference, following up the last chapter, and robust clustering.
The second extension we propose deals with the case where we observe n i.i.d. measures $X_1, \ldots, X_n$ drawn from X rather than n i.i.d. points. In this framework, quantizing the mean measure $\mathbb{E}(X)$ with k points provides a way to convert the measure sample into a $\mathbb{R}^k$-valued sample, by roughly encoding how much mass does a measure spread around every code point. After investigating the properties of a classical quantization algorithm (Mac Queen algorithm) adapted to the measure sample case, we provide some insights on the ability of our embedding strategy to discriminate between different classes of measures. A theoretical case study is conducted in clustering a measure sample drawn form a mixture of persistence diagram. At last, some experimental results in graph classification tasks are exposed.

3.1 Robust Bregman quantization

3.1.1 Outliers modeling and robustness

Following up the example given in Section 2.3.4, where 80 of the 280 samples were drawn uniformly on ambient cubes, some noisy situations involve outlier points. The example given in Section 2.3.4 is a particular instance of the so-called Huber contamination model, defined by

$$Q \sim (1-\alpha)P + \alpha N, \tag{3.1}$$

where P is the signal distribution (supported on a compact set K in our topological inference framework) and N is a noise distribution. In this model, a portion α of the sample points are sampled according to a noise distribution N, that are called *outliers*. Up to our knowledge, this is the first introduced model that accounts for corrupted samples (see, e.g., Huber [1964, 1965]). In some sense, robustness to Huber's contamination model is already alleged for the DTM and k-PDTM. Indeed, if Q follows (3.1), then $W_2(P,Q) \leq \sqrt{\alpha} W_2(P,N)$, so that Theorem 2.12 and Proposition 2.17 may apply whenever $W_2(P,N)$ is controlled. However, such stability results do not allow to identify and remove outlier points, that could be of interest to carry further topological or geometric inference.
More concerning, Wasserstein stability of objects does not allow to face more mischievous contamination models, such as the adversarial contamination model, where a portion ε of the sample can have arbitrary values (not necessarily i.i.d. with respect to a noise distribution). To be more precise, the adversarial noise model assumes that we observe

$$\mathbb{X}_{n+s}^c = \mathbb{X}_n \cup \{x_1, \dots, x_s\},$$

where $\mathbb{X}_n$ originates from the signal distribution (possibly with Wasserstein noise), and $\{x_1, \dots, x_s\}$ are points in $\mathbb{R}^D$. This corresponds to the situation where a portion $s/(n+s)$ of the sample is arbitrarily corrupted. Since Wasserstein stability is no more relevant in this framework, robustness of statistical procedures may be alleged using the so-called *Breakdown Point* Donoho and Huber [1983].

Definition 3.1 :

Let $\hat{T}$ be a vector-valued estimator. The *Finite Sample Breakdown Point* of $\hat{T}$ is

$$\inf \left\{ \frac{s}{n+s} \mid \sup_{x_1,\dots,x_s} \left\| \hat{T}(\mathbb{X}_{n+s}^c) \right\| = +\infty \right\}.$$

Intuitively speaking, the FSBP (Finite Sample Breakdown Point) is the smallest proportion of adversarial noise in a sample that can render an estimator arbitrarily bad, for a fixed signal sample. For instance, in a central tendency estimation framework, the empirical mean has a FSBP of $1/(n+1)$ (one corrupted data point is enough to drive the empirical mean at infinity), whereas the empirical median has breakdown point $(n+1)/(2n)$. Note that FSBP do not assess quality of estimation (for instance constant estimators have infinite FSBP). Therefore, they should be combined with other results (such as oracle bounds) to evaluate the overall performance of an estimator in a noisy setting.
Let us mention that other contamination models and measures of robustness exist. Concerning the contamination models, the Huber noise model might be considered as the easier to deal with, whereas the adversarial one is the most difficult. A panoramic view of different and intermediate contamination models may be found in [Bateni and Dalalyan, 2020, Section 2]. Other methods to seize robustness of estimators are exposed in [Maronna et al., 2019, Section 3] or Bateni and Dalalyan [2020].

3.1.2 Trimmed Bregman quantization

To identify outlier sample points based on a distance to compact estimator $\hat{\mathrm{d}}$, a standard way could be to compute $\{\hat{\mathrm{d}}(\mathbb{X}_{n+s}^c)\}$ and to discard the points corresponding to the most extreme values. Though this approach seems to work well on the example given in Section 2.3.4 (see Figure 2.3), it strongly relies on the fact that $\hat{\mathrm{d}}$ is close to d_K beforehand, that is not a realistic assumption. For instance, in our k-PDTM framework,

if the s outliers are concentrated at a single location, it is likely that one code point of an optimal codebook for the k-PDTM would account for this location, rendering the aforementioned strategy irrelevant.
To cope with the adjunction of a portion of possibly structured noise (whose proportion α is known beforehand), in a mean estimation framework, one can resort to *trimmed mean*, that is

$$\hat{m}_\alpha \in \arg\min_m \min_{\{Y \subset \mathbb{X}_{n+s}^c, |Y| \geq (1-\alpha)|\mathbb{X}_{n+s}|\}} \frac{1}{|Y|} \sum_{y \in Y} \|y - m\|^2.$$

In other words, the trimmed mean is the "best" possible mean one could get from a $1-\alpha$ sub-sample, quality of these subsamples means being seized by their corresponding L_2 inertias. From a theoretical viewpoint, this amounts to optimize $(Q_\alpha, m) \mapsto Q_\alpha(du)\|u-m\|^2$, where $Q_\alpha \in \mathcal{P}_{1-\alpha}(Q)$ ranges into the set of $(1-\alpha)$ sub-measures of Q as defined in (2.2). Robustness of the trimmed mean may be seized in terms of FSBP: indeed, for a trimming level α, its FSBP is roughly α (see, e.g. [Maronna et al., 2019, Section 3.2.5]). Note that other robust mean estimator may be conceived with better theoretical properties, based on PAC-Bayesian or MOM techniques for instance (Catoni and Giulini [2018], Brownlees et al. [2015], Lecué and Lerasle [2020]). However, this trimming approach also allows to simply identify outliers (the α portion of mass that is discarded), and easily adapts to every M-estimation procedure.
In the quantization framework, robustification of the k-means scheme has been proposed in Cuesta-Albertos et al. [1997], via the minimization of the so-called *trimmed k-variation*

$$(Q_\beta, \mathbf{c}) \mapsto Q_\beta(du) \min_{j=1,\dots,k} \|u - c_j\|^2,$$

where $Q_\beta \in \mathcal{P}_{(1-\alpha)+}(Q)$ that is the set of set of s-sub-measures of Q, for $s \geq 1-\alpha$,

$$\mathcal{P}_{(1-\alpha)+}(Q) = \bigcup_{s \geq (1-\alpha)} \mathcal{P}_s(Q). \tag{3.2}$$

This heuristics results in a sample-based robust quantization procedure called *trimmed k-means*.
To encompass the special case of the k-PDTM, this method can be adapted to quantization with generic Bregman divergences. Let us recall here the definition informally given in Section 2.3.4.

Definition 3.2 :

> Let ϕ be a $\mathcal{C}^1$ convex real-valued function defined on a convex set $\Omega \subset \mathbb{R}^D$. The Bregman divergence d_ϕ is defined, for all $x, y \in \Omega$, by
>
> $$\mathrm{d}_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla_y \phi, x - y \rangle .$$

The choice $\phi(u) = \|u\|^2$ corresponds to the trimmed k-means of Cuesta-Albertos et al. [1997], whereas $\phi(u) = \|u\|^2 - \mathrm{d}^2_{Q,h}(u)$ corresponds to the k-PDTM case (note that in this case ϕ is $\mathcal{C}^1$ Q a.s. whenever $Q(du)\|u\|^2 < +\infty$, according to [Chazal et al., 2011, Corollary 3.7]). Some Bregman divergences are of particular interest in exponential mixture models, as will be detailed in Section 3.1.6. A key property of Bregman divergences is that means are minimizers of Bregman inertias.

Proposition 3.3 : [Banerjee et al., 2005b, Proposition 1]

Let Q be such that $Q(du)\|u\| < \infty$, $Q(du)u \in \mathring{\Omega}$, and $Q(du)\phi(u) < \infty$. Then, for any $x \in \Omega$,

$$Q(du)\mathrm{d}_\phi(u,x) = Q(du)\mathrm{d}_\phi(u, Q(du)u) + \mathrm{d}_\phi(Q(du)u, x).$$

Uniqueness of Bregman inertia minimizers follows from Proposition 3.3 whenever ϕ is strictly convex. Interestingly, Bregman divergences are in fact characterized by this property (see, e.g., [Banerjee et al., 2005a, Theorem 3]). Slightly anticipating, this property allows for iterative algorithm to search for local minimizers, by iteratively replacing the center of a code cell V_j by its mean, as described in Banerjee et al. [2005b]. Similarly to the trimmed k-means case, we intend to minimize

$$R(Q_\alpha, \mathbf{c}) = Q_\alpha(du)\mathrm{d}_\phi(u, \mathbf{c}), \tag{3.3}$$

where $\mathbf{c} \in (\mathbb{R}^D)^k$, $\mathrm{d}_\phi(u, \mathbf{c}) = \min_{j=1,\dots,k} \mathrm{d}_\phi(u, c_j)$, and Q_α ranges in $\mathcal{P}_{(1-\alpha)_+}(Q)$. Structural results on minimizers of (3.3) may be derived, in a similiar way to Gordaliza [1991], Cuesta-Albertos et al. [1997]. For $\mathbf{c} \in (\mathbb{R}^D)^k$ and $r > 0$, we denote by $\mathrm{B}_\phi(\mathbf{c}, r)$ (resp. $\bar{\mathrm{B}}_\phi(\mathbf{c}, r)$) the Bregman ball $\{x \mid \sqrt{\mathrm{d}_\phi(x, \mathbf{c})} < r\}$ (resp. $\leq$), and by $r_\alpha(\mathbf{c})$ the radius r such that

$$Q(\mathrm{B}_\phi(\mathbf{c}, r_\alpha(\mathbf{c})) \leq 1 - \alpha \leq Q(\bar{\mathrm{B}}_\phi(\mathbf{c}, r_\alpha(\mathbf{c})).$$

Next, for $\mathbf{c} \in (\mathbb{R}^D)^k$ and $\alpha \in [0,1]$, the following subset of $(1-\alpha)$-submeasures of Q is of particular interest:

$$\mathcal{P}_{\mathbf{c},\alpha}(Q) = \left\{ \tilde{Q} \in \mathcal{P}_{(1-\alpha)}(Q) \mid \tilde{Q}_{|\mathrm{B}_\phi(\mathbf{c}, r_\alpha(\mathbf{c}))} = Q_{|\mathrm{B}_\phi(\mathbf{c}, r_\alpha(\mathbf{c}))} \quad \text{and} \quad \tilde{Q}_{|\bar{\mathrm{B}}_\phi(\mathbf{c}, r_\alpha(\mathbf{c}))} = 1 - \alpha \right\},$$

that is the set of $(1-\alpha)$-submeasures that minimize $Q_\alpha \mapsto Q_\alpha(du)\mathrm{d}_\phi(u, \mathbf{c})$. Equipped with these definitions, we may define the α-trimmed distortion of a codebook $\mathbf{c}$ via

$$R_\alpha(\mathbf{c}) = (1-\alpha)Q_{\mathbf{c},\alpha}(du)\mathrm{d}_\phi(x, \mathbf{c}) = (1-\alpha)R(Q_{\mathbf{c},\alpha}, \mathbf{c}),$$

for $Q_{\mathbf{c},\alpha} \in \mathcal{P}_{\mathbf{c},\alpha}(Q)$. In the case where $Q(\bar{\partial}\mathrm{B}_\phi(\mathbf{c}, r_\alpha(\mathbf{c}))) = 0$ (for instance whenever Q has a density), $R_\alpha(\mathbf{c})$ is just $Q(du)\mathrm{d}_\phi(u, \mathbf{c})\mathbb{1}_{\mathrm{B}_\phi(\mathbf{c}, r_\alpha(\mathbf{c}))}(u)$. The following result ensures that minimizing (3.3) is equivalent to minimize the α-trimmed distortion.

Proposition 3.4 :

Let $\alpha \in [0,1]$. Then

$$(1-\alpha) \min_{Q_\alpha \in \mathcal{P}_{(1-\alpha)_+}(Q), \mathbf{c} \in (\mathbb{R}^D)^k} R(Q_\alpha, \mathbf{c}) = \min_{\mathbf{c} \in (\mathbb{R}^D)^k} R_\alpha(\mathbf{c}).$$

Proposition 3.4 generalizes [Cuesta-Albertos et al., 1997, Proposition 2.3], and ensures that optimal trimming sets are in fact Bregman balls (up to boundary effects). Thus, seeking for both optimal codebooks and trimming sets boils down to minimize the α-trimmed distortion, that may be carried out using an iterative scheme using Proposition 3.3. For the sake of completeness, it remains to ensure that optimal codebooks for the trimmed distortion do exist.

THEOREM 3.5 :

Let $\alpha \in]0,1[$, and assume that $\overline{\text{conv}\,(\text{Supp}(Q))} \subset \mathring{\Omega}$. If $Q(du)\|u\| < +\infty$, then the set $\arg\min_{\mathbf{c}\in(\mathbb{R}^D)^k} R_\alpha(\mathbf{c})$ is not empty.

This result generalizes [Cuesta-Albertos et al., 1997, Theorem 3.1] as well as [Fischer, 2010, Theorem 3.1]. Note that the assumption $Q(du)\|u\| < +\infty$ is slightly milder than the usual $Q(du)\|u\|^2 < +\infty$ that is required for the existence of optimal codebooks in the usual squared-distance quantization framework (see, e.g., Pollard [1981]). From now on we denote by $\mathbf{c}^*_\alpha$ a minimizer of R_α, and intend to approximate such an optimal trimmed codebook from sample.

3.1.3 Iterative algorithm

The main idea behind the iterative algorithm that is exposed below is based on the following Proposition 3.7. To formalize the assignment of one point x to a code point, we introduce the so-called *code cells*.

DEFINITION 3.6 :

Let $\mathbf{c} \in (\mathbb{R}^D)^k$. The subsets $C_1(\mathbf{c}), \ldots, C_k(\mathbf{c})$ are *code cells* for $\mathbf{c}$ if

(i) $(C_1(\mathbf{c}), \ldots, C_k(\mathbf{c}))$ is a partition of $\mathbb{R}^D$;

(ii) For all $j \in [\![1, k]\!]$,

$$\mathring{V}_j(\mathbf{c}) \subset C_j(\mathbf{c}) \subset V_j(\mathbf{c}),$$

where $V_j(\mathbf{c})$ is the Voronoi cell associated with c_j.

The notion of code cells allows to switch from codebooks to quantizers, that is maps from $\mathbb{R}^D$ to a subset of k points, by defining the quantizers $q_\mathbf{c}$ that map every $C_j(\mathbf{c})$ onto c_j. It is designed to bypass the assignment of Voronoi boundaries $\partial(V_j(\mathbf{c}) \cap V_i(\mathbf{c}))$. Recall that according to [Graf et al., 2007, Proposition 1], $Q(V_i(\mathbf{c}^*) \cap V_j(\mathbf{c}^*)) = 0$, so that the assignment of boundaries is somehow artificial for optimal codebooks. Since there exists an infinite number of code cell choices associated with a codebook, to fix the ideas we choose to assign boundaries with the lexicographic order, that is

$$C_j(\mathbf{c}) = V_j(\mathbf{c}) \cap \left(\bigcap_{i<j} V_i(\mathbf{c}) \right)^c,$$

so that code cells associated with $\mathbf{c}$ are well-defined. Code cells being defined, we are in position to state the key result of this section.

PROPOSITION 3.7 :

Let $\mathbf{c} \in \Omega^{(k)}$ and $\tilde{Q}_\mathbf{c} \in \mathcal{P}_{\mathbf{c},\alpha}(Q)$. Assume that for all $i \in [\![1, k]\!]$, $\tilde{Q}_\mathbf{c}(C_i(\mathbf{c})) > 0$, and denote by $\mathbf{m}$ the codebook of the local means of $\tilde{Q}_\mathbf{c}$. In other words, $m_i = \tilde{Q}_\mathbf{c}(u\mathbb{1}_{C_i(\mathbf{c})}(u))/\tilde{Q}_\mathbf{c}(C_i(\mathbf{c}))$. Then

$$R_\alpha(\mathbf{c}) \geq R_\alpha(\mathbf{m}),$$

with equality if and only if for all i in $[\![1, k]\!]$, $\mathrm{d}_\phi(m_i, c_i) = 0$.

Proposition 3.7 is a straightforward consequence of Proposition 3.3, that emphasizes the key property that Bregman divergences are minimized by expectations (this is not the case for the L_1 distance for instance). Note that whenever ϕ is strictly convex, $\mathrm{d}_\phi(m_i, c_i) = 0$ if and only if $c_i = m_i$, so that a codepoint modification entails a decrease of the risk function in this case. In the DTM case, where $\psi(u) = \|u\|^2 - \mathrm{d}^2_{Q,h}(u)$, it holds $d_\phi(m_i, c_i) = 0$ if and only if $Q_{c_i,h} \in \mathcal{P}_{m_i,h}(Q)$ (see for instance [Brécheteau and Levrard, 2020, Proposition 9]). In particular, if $Q = Q_n$ and $q = nh$, equality holds if and only if c_i and m_i have the same q nearest neighbors.
In line with Banerjee et al. [2005b] for the non-trimmed case, Proposition 3.7 provides an iterative scheme to minimize R_α The algorithm we introduce is inspired by the trimmed version of Lloyd's algorithm Cuesta-Albertos et al. [1997], and is also a generalization of the Bregman clustering algorithm [Banerjee et al., 2005b, Algorithm 1]. Recall that $\{X_1, \dots, X_n\} = \mathbb{X}_n$, and that Q_n denotes the empirical distribution $\sum_{i=1}^n \delta_{x_i}$. Further, we assume that α satisfies $(1 - \alpha) = \frac{q}{n}$ for some positive integer q.

Algorithm 1 : Bregman trimmed k-means.

- **Input:** $\{X_1, \dots, X_n\} = \mathbb{X}_n$, q, k.
- **Initialization:** Sample c_1, c_2,$\dots$ c_k from $\mathbb{X}_n$ without replacement, $\mathbf{c}^{(0)} \leftarrow (c_1, \dots, c_k)$.
- **Iterations:** Repeat until stabilization of $R^{(t)}_{n,q}$.
  - $NN^{(t)}_q \leftarrow$ indices of the q smallest values of $d_\phi(x, \mathbf{c}^{(t-1)})$, $x \in \mathbb{X}_n$.
  - For $j = 1, \dots, k$, $C^{(t)}_{j,q} \leftarrow C_j(\mathbf{c}^{(t-1)}) \cap NN^{(t)}_q$.
  - Compute $R^{(t-1)}_{n,q} = \frac{1}{q} \sum_{j=1}^k \sum_{x \in C^{(t)}_{j,q}} \mathrm{d}_\phi(x, c^{(t-1)}_j)$.
  - For $j = 1, \dots, k$, $c^{(t)}_j \leftarrow \dfrac{\sum_{x \in C^{(t)}_{j,q}} x}{\left|C^{(t)}_{j,q}\right|}$.
- **Output:** $\mathbf{c}^{(t)}$, $C^{(t)}_{1,q}, \dots, C^{(t)}_{k,q}$.

As a remark, whenever ϕ is strictly convex, the stopping criterion in Algorithm 1 can be replaced with the stabilization of $\mathbf{c}^{(t)}$, and the computation of $R^{(t)}_{n,q}$ can be spared. As for every EM-type algorithm, initialization may be crucial. Section 3.2.1 gives some insight of what is a good initialization and on the probability that such a good initialization is obtained by random sampling of $\mathbb{X}_n$, in a simpler case. In practice, several random starts will be proceeded. More sophisticated strategies, such as `k-means ++` Arthur and Vassilvitskii [2007], could be an efficient way to address the initialization issue.
An easy consequence of Proposition 3.7 for the empirical measure Q_n associated with $\mathbb{X}_n$ is the following. For short we denote by $R_{n,\alpha}$ the trimmed distortion associated with Q_n.

Proposition 3.8 :

> *Algorithm 1 converges to a local minimum of the function $R_{n,\alpha}$.*

It is worth mentioning that in full generality the output of Algorithm 1 is not a global minimizer of $R_{n,\alpha}$. However, it is likely that suitable clusterability assumptions as in Kumar and Kannan [2010], Tang and Monteleoni [2016], Levrard [2018] would lead to further guarantees on such an output. Section 3.2.1 gives an instance where such results can be obtained, in a simpler case though.

3.1.4 Theoretical guarantees

Theoretical guarantees for our sample-based robust quantization scheme are given under slightly stronger conditions on the underlying Bregman divergence and sampling distribution Q. In this section we further assume that

(i) ϕ is strictly convex;

(ii) ϕ is $\mathcal{C}^2$;

(iii) $\overline{\mathrm{conv}(\mathrm{Supp}(Q)} \subset \mathring{\Omega}$.

Note that these assumptions are satisfied for $\|u\|^2 - \mathrm{d}^2_{Q,h}(u)$ whenever Q has a continuous density on $\mathbb{R}^D$, using for instance [Chazal et al., 2011, Proposition 3.6] and Baddeley [1977].

3.1.4.1 Convergence of empirically optimal codebooks

Recall that $R_{n,\alpha}$ denotes the α-trimmed empirical distortion (associated with the empirical distribution Q_n). We let $\hat{\mathbf{c}}_{n,\alpha}$ denote an empirically optimal trimmed k-points codebook, that is a minimizer of $R_{n,\alpha}$, and investigate the generalization properties of $\hat{\mathbf{c}}_{n,\alpha}$ (that is how it compares with optimal codebooks for R_α). We begin with an asymptotic convergence result.

Theorem 3.9 :

Assume that Q is absolutely continuous with respect to the Lebesgue measure and satisfies $Q\|u\|^p < \infty$ for some $p > 2$, then there exists $\mathbf{c}^*_\alpha$ an optimal codebook such that

$$\lim_{n\to+\infty} R_{n,\alpha}(\hat{\mathbf{c}}_{n,\alpha}) = R_\alpha(\mathbf{c}^*_\alpha) \ a.e..$$

Moreover, up to extracting a subsequence, we have

$$\lim_{n\to+\infty} D(\hat{\mathbf{c}}_{n,\alpha}, \mathbf{c}^*_\alpha) = 0 \ a.e.,$$

where $D(\mathbf{c}, \mathbf{c}') = \min_{\sigma\in\Sigma_k} \max_{i\in[\![1,k]\!]} |c_i - c'_{\sigma(i)}|$ and Σ_k denotes the set of all permutations of $[\![1, k]\!]$. At last, if $\mathbf{c}^*_\alpha$ is unique, then $\lim_{n\to+\infty} D(\hat{\mathbf{c}}_{n,\alpha}, \mathbf{c}^*_\alpha) = 0$ a.e. (without taking a subsequence).

This result generalizes [Cuesta-Albertos et al., 1997, Theorem 3.4], ensuring that empirically optimal trimmed k-points codebooks converge to optimal trimmed k-points codebooks. In the Huber contamination model $Q \sim (1-\gamma)P + \gamma N$, ensuring that optimal trimmed codebooks do not account for noise seems possible if the noise distribution is far enough from P. Example 3.10 gives an instance of such a result.

Example 3.10. We let $\phi(u) = \|u\|^2$. Let $P \sim \mathcal{U}(-1, 1)$, and $Q \sim (1-\gamma)P + \gamma\delta_N$, where $N \geq 5$. Assume that $0 \leq \gamma \leq \alpha \leq 0.18$, and $k = 2$. Then $\mathbf{c}^*_\alpha$ is an optimal two-point codebook for $\mathcal{U}\left(-\frac{1-\alpha}{1-\gamma}, \frac{1-\alpha}{1-\gamma}\right)$.

Proof of Example 3.10. Let $\mathbf{c} = (-\frac{1}{2}, \frac{1}{2})$. Then straightforward computation leads to $R_\alpha(\mathbf{c}) \leq \frac{1-\gamma}{12}$.

On the other hand, assume that $\{N\} \in \bar{\mathrm{B}}(\mathbf{c}, r_\alpha(\mathbf{c}))$, and $c_1 \leq c_2$. Then $c_2 \geq \frac{N-1}{2} \geq 2$, so that

$$\begin{aligned} R_\alpha(\mathbf{c}) &\geq (1-\gamma) P(du) \|u - c_1\|^2 \mathbb{1}_{\bar{\mathrm{B}}(\mathbf{c}, r_\alpha(\mathbf{c}))}(u) \\ &\geq (1-\gamma) \int_0^{\frac{1-(\gamma+\alpha)}{1-\gamma}} u^2 du \\ &\geq \frac{(1-(\gamma+\alpha))^2}{3(1-\gamma)^3} > \frac{1-\gamma}{12} \end{aligned}$$

whenever $0 \leq \gamma \leq \alpha \leq 0.18$. Thus, for any optimal codebook $\mathbf{c}_\alpha^*$, $\{N\} \notin \bar{\mathrm{B}}(\mathbf{c}^*, r_\alpha(\mathbf{c}^*))$, hence $\mathbf{c}_\alpha^*$ is an optimal codebook for $(1-\gamma)\mathcal{U}(-1+, 1)$ at trim level $\alpha - \gamma$. Straightforward calculation gives that $\mathbf{c}_\alpha^*$ is then an optimal codebook for $\mathcal{U}\left(-\frac{1-\alpha}{1-\gamma}, \frac{1-\alpha}{1-\gamma}\right)$. □

As exposed above, providing guarantees on trimmed optimal codebooks in the Huber contamination model heavily depends on the structure of P and N, thus providing a general result seems hard. For instance, if the noise distribution is close to the support of P, then the benefit of trimming this noise might be smaller than trimming P. As well, the benefit of noise trimming also depends on the portion γ: for large γ's, quantizing N and trimming P could yield better trimmed distortions whenever N is more concentrated than P. In this case, the distinction between noise and signal seems intuitively less relevant, and cannot be carried out via trimmed L^2 criterions. More insights on the structures of P that guarantee the non-degeneracy of optimal codebooks are given in the following Section 3.1.4.2.
In addition to Theorem 3.9, non-asymptotic bounds on the excess trimmed distortion of an empirically optimal codebook may be derived.

THEOREM 3.11 :

Assume that $Q\|u\|^p < \infty$, where $p \geq 2$. Further, if $R_{k,\alpha}^*$ denotes the α-trimmed optimal distortion with k points, assume that $R_{k-1,\alpha}^* - R_{k,\alpha}^* > 0$. Then, for n large enough, with probability larger than $1 - n^{-\frac{p}{2}} - 2e^{-x}$, we have

$$R_\alpha(\hat{\mathbf{c}}_{n,\alpha}) - R_\alpha(\mathbf{c}_\alpha^*) \leq \frac{C_Q}{\sqrt{n}}(1 + \sqrt{x}).$$

The requirement $R_{k-1,\alpha}^* - R_{k,\alpha}^* > 0$ ensures that optimal codebooks will not have empty cells. Note that if $R_{k-1,\alpha}^* - R_{k,\alpha}^* = 0$, then there exists a subset A of $\mathbb{R}^D$ satisfying $Q(A) \geq 1 - \alpha$ and such that the restriction of Q to A is supported by at most $k-1$ points, that allows optimal k-points codebooks with at least one empty cell. It is worth mentioning that Theorem 3.11 does not require a unique trimmed optimal codebook, and only requires an order 2 moment condition for $\hat{\mathbf{c}}_{n,\alpha}$ to achieve a sub-Gaussian rate in terms of trimmed distortion. This condition is in line with the order 2 moment condition required in Brownlees et al. [2015] for a robustified estimator of $\mathbf{c}^*$ to achieve similar guarantees, as well as the finite-variance condition required in Catoni and Giulini [2018] in a mean estimation framework.
As well, Theorem 3.11 can be used to non-asymptotically assess that the outliers are removed by empirically optimal codebooks, in some particular instances of Huber's contamination model. Indeed, following up Example 3.10, in this case not trimming the noise located at $\{N\}$ incurs a lower bound on the distortion, that may be combined with Theorem 3.11 to ensure that it is impossible for an empirically optimal codebook with high probability. It falls under the intuition that a key quantity to guarantee noise

trimming is $R^*_{k-1,\alpha} - R^*_{k,\alpha}$, that seizes how much costs the removal of one code point to the signal. If this cost is negligible compared to not quantizing the noise, then it is likely that optimal codebooks will be degenerate with respect to the noise. This heuristic is at the core of the robustness properties exposed below.

3.1.4.2 Breakdown point

We recall that the corrupted sample (with s outliers) is denoted by $\mathbb{X}^c_{n+s} = \mathbb{X}_n \cup \{x_1, \dots, x_s\}$, where $\{x_1, \dots, x_s\}$ are (possibly adversarial) outliers and $\mathbb{X}_n$ is drawn from Q. Further, we denote by Q_{n+s} the empirical distribution associated with $\mathbb{X}^c_{n+s}$ and by $\hat{\mathbf{c}}_{n+s,\alpha}$ a minimizer of $\mathbf{c} \mapsto R_{n+s,\alpha}(\mathbf{c})$. We intend to seize the robustness of our robust Bregman quantization scheme in terms of Breakdown point (Definition 3.1), that is the smallest proportion of outliers that can drive $\|\hat{\mathbf{c}}_{n+s,\alpha}\|$ to infinity.
To give an intuition, the standard mean (minimizer of the 0-trimmed empirical distortion for $k = 1$, $\phi(u) = \|u\|^2$) has breakdown point $\frac{1}{n+1}$, whereas α-trimmed means have breakdown point roughly α (see, e.g., [Maronna et al., 2019, Section 3.2.5]). According to [Vandev, 1993, Theorem 1], this is also the case whenever ϕ is strictly convex (but still $k = 1$). In the case $k > 1$, as noticed in Cuesta-Albertos et al. [1997] for trimmed k-means, the breakdown point may be much smaller than α. As mentioned above, if an α-trimmed optimal codebook has a too small cluster, then adding an adversarial cluster with greater weight might switch the roles between noise and signal, resulting in an α-trimmed codebook that allocates one point to the adversarial cluster and trims the too small optimal cluster. To quantify this intuition, we introduce the following discernability factor B_α.

Definition 3.12 :

Let $\alpha \in]0, 1[$, and, for $b \leq (1 - \alpha)$, denote by $\alpha_b^- = (\alpha - b)/(1 - b)$, $\alpha_b^+ = \alpha/(1 - b)$. The discernability factor B_α is defined as

$$B_\alpha = \sup \left\{ b \geq 0 \mid b \leq \alpha \wedge (1 - \alpha) \text{ and } \min_{j \in [\![2,k]\!]} R^*_{j-1,\alpha_b^+} - R^*_{j,\alpha_b^-} > 0 \right\}.$$

In fact, $(\alpha^+_{B_\alpha} - \alpha) = \alpha B_\alpha/(1 - B_\alpha)$ is the portion of the $(1 - \alpha)$ mass in an optimal k-points trimming set that may be considered as noise by an optimal $k - 1$-points with trim level $\alpha^+_{B_\alpha}$. As exposed in the following proposition, B_α is related to the minimum cluster weight of optimal α-trimmed codebooks.

Proposition 3.13 :

Under the assumptions of Section 3.1.4, if $R^*_{k-1,\alpha} - R^*_{k,\alpha} > 0$, then $B_\alpha > 0$. Moreover, for any $j \in [\![1, k]\!]$, if $\mathbf{c}^{*,(j)}$ is an optimal j-points α-trimmed codebook and $p_{j,\alpha} = (1 - \alpha) \min_{p \in [\![1,j]\!]} \tilde{Q}_{\mathbf{c}^{*,(j)}}\left(C_p(\mathbf{c}^{*,(j)})\right)$, with $\tilde{Q}_{\mathbf{c}^{*,(j)}} \in \mathcal{P}_{\mathbf{c}^{*,(j)},1-\alpha}(Q)$, then

$$B_\alpha \leq \frac{p_{j,\alpha}}{\alpha + p_{j,\alpha}}.$$

Proposition 3.13 emphasizes that, for a fixed trim level α, the discernability factor B_α may become arbitrarily small whenever the minimal weight of an optimal α-trimmed cell decreases. Theorem 3.14 below makes connection between this discernability factor and robustness properties of optimal k-points α-trimmed codebook, stated in terms of Bregman radius.

THEOREM 3.14 :

For $\ell \geq 1$, let $R^*_{\ell,\alpha}$ denote the ℓ-points α-trimmed optimal distortion. Assume that $Q\|u\|^p < +\infty$, for some $p \geq 2$. Moreover, assume that $R^*_{k-1,\alpha} - R^*_{k,\alpha} > 0$. Let $b < B_\alpha$, and assume that $s/(n+s) \leq b$. Then, for n large enough, with probability larger than $1 - n^{-\frac{p}{2}}$,

$$\max_{j \in [\![1,k]\!]} \mathrm{d}_\phi\left(\mathrm{B}(0, C_{Q,b}), \hat{c}_{n+s,\alpha,j}\right) \leq K_{Q,b},$$

where $C_{Q,b}$ and $K_{Q,b}$ do not depend on n nor s.

Theorem 3.14 guarantees that the proposed trimming procedure is robust in terms of Bregman divergence, that is, the corrupted empirical distortion minimizer belongs to some closed Bregman ball, provided the proportion of noise is smaller than the discernability factor introduced in Definition 3.12. Unfortunately Bregman balls might not be compact sets if $c \mapsto \mathrm{d}_\phi(x, c)$ is not a proper map. For instance, with $\phi(x) = e^x$ and $\Omega = \mathbb{R}$, we have $]-\infty, 0] \subset \{c \mid \mathrm{d}_\phi(0, c) \leq 1\}$. In the proper map case, Theorem 3.14 entails that the FSBP is larger than B_α, with high probability, for n large enough. In the other case, Corollary 3.15 below ensures that this breakdown point is positive, provided that $p > 2$. We denote by $\hat{B}_{n,\alpha}$ the FSBP of our trimmed Bregman quantization estimator.

COROLLARY 3.15 :

Assume that $Q\|u\|^p < +\infty$, for $p > 2$. Under the assumptions of Theorem 3.14, there exists $c > 0$ such that, almost surely, for n large enough, $\widehat{BP}_{n,\alpha} \geq c$.

In addition, if, for every $x \in \Omega$, $c \mapsto \mathrm{d}_\phi(x, c)$ is a proper map, then almost surely, for n large enough $\widehat{BP}_{n,\alpha} \geq B_\alpha$.

Corollary 3.15 guarantees that our trimmed Bregman quantization procedure is asymptotically robust in the usual sense to a certain proportion of adversarial noise, contrary to plain Bregman quantization whose FSBP is $1/(n+1)$. However this unknown authorized proportion depends on both the choice of Bregman divergence and the discernability factor B_α. In the proper map case, the FSBP is larger than B_α. Note that for $x \in \Omega$, $c \mapsto \mathrm{d}_\phi(x, c)$ is proper whenever ϕ is strictly convex, that is the case for trimmed k-means Cuesta-Albertos et al. [1997]. For this particular Bregman divergence, the result of Corollary 3.15 is provably tight.

Example 3.16 *: [Brécheteau et al., 2021, Example 17].* Let $\phi_1 = \|.\|^2$, $\phi_2 = \exp(-.)$, $\Omega = \mathbb{R}$, $P = (1-p)\delta_{-1} + p\delta_1$, with $p \leq 1/2$. Then, for $\phi = \phi_j$, $j \in \{1, 2\}$, $k = 2$ and $p > \alpha$, we have $B_\alpha = \frac{p-\alpha}{p} \wedge \alpha$. Let $Q_{\gamma,N} = (1-\gamma)P + \gamma\delta_N$. The following holds.

- If $(1+p)(1-\alpha) > 1$, $B_\alpha = \alpha$, and for every $\gamma > \alpha$, any sequence of optimal 2-points h-trimmed codebook $\mathbf{c}^*_2(Q_{\gamma,N})$ for $Q_{\gamma,N}$ satisfies

$$\lim_{N \to +\infty} \|\mathbf{c}^*_2(Q_{\gamma,N})\| = +\infty.$$

- If $(1+p)(1-\alpha) \leq 1$, then $B_\alpha = \frac{p-\alpha}{p}$, and, for $\gamma = B_\alpha$, $(-1, N)$ is an optimal 2-points α-trimmed codebook for $Q_{\gamma,N}$.

Note that upper bounds on the FSBP when $n \to +\infty$ may be derived for Example 3.16 using standard deviation bounds. Example 3.16 illustrates the two situations that can be encountered when some adversarial noise is added, depending on the balance between trim level and smallest optimal code cell. If the trim level is small enough compared to

the smallest mass of an optimal cell (first case), then the breakdown point is simply α, that is the amount of points that can be trimmed. This corresponds to the breakdown point of the trimmed mean (see, e.g., Vandev [1993]). When the trim level becomes large compared to the smallest mass of an optimal cell (second case), optimal codebooks for the perturbed distribution can be codebooks that allocate one point to the noise and trim the small optimal cell, leading to a breakdown point possibly smaller than α. This corresponds to the situation exposed in Proposition 3.13. In both cases, the breakdown point is smaller than B_α, thus, according to Corollary 3.15, it is equal to B_α.
As mentioned in Cuesta-Albertos et al. [1997] for the trimmed k-means, in practice, breakdown point and choice of the correct number of cells k are closely related questions, that depend on the structure of the noise. Namely, in an adversarial framework, the proportion α_0 of noise that can be satisfactorily processed via a robust quantization scheme with trimming level α_0 must be smaller than $1/k$, according to Corollary 3.15, Example 3.16 and Proposition 3.13. In practice, if the noise is less concentrated than the signal, for instance in a Huber contamination model with ambient noise and signal on a lower dimensional structure, it is likely that larger proportion of noises may be successfully managed. This phenomenon is empirically illustrated in Section 3.1.5. Providing general bounds on authorized proportions of noise in Huber's contamination model remains an open question.
At last, in a clustering framework (that is in the case where the signal is naturally organized around k_0 poles), some heuristic may be given to empirically choose both the number of clusters k and the trim level α. This point is illustrated in Section 3.1.6, where the correct number of clusters depends on what is considered as noise.

3.1.5 Outliers detection (and topological inference) with robust k-PDTM

We come back to the experiment of Section 2.3.4. Recall that 200 points are uniformly drawn on the infinity symbol K, with an additional additive Gaussian noise, with standard deviation $\sigma = 0.02$. Then 80 outliers are added: 40 points are uniformly drawn on the rectangle $[-2, 5] \times [-2, 2]$, and 40 points on the rectangle $[-4, 7] \times [-4, 4]$. This results in a proportion $\alpha_0 = 2/7$ of outliers, for a corrupted sample $\mathbb{X}_{n+s}^c$ with 280 points. First we illustrate the naive approach that consists in removing the $(n+s)\alpha_0$ points that corresponds to the largest values of $\hat{\mathrm{d}}$, for estimates $\hat{\mathrm{d}}$ of d_K given by the k-means, q-witness distance, and k-PDTM. This general scheme is referred to as *truncation*, resulting in *truncated* k-means, *truncated* q-witnessed and *truncated* k-PDTM. As in Section 2.3.4, we choose $q = 10$ and $k = 50$. Once outliers removed, we remove code points corresponding to empty cells and compute the persistence diagram associated with the resulting (truncated) distance estimate.
Truncated k-PDTM and q-witnessed distance seem to correctly grid K and identify outliers. Compared with Figure 2.5, the removed code points correspond to code points with high weight ω_Q^2, that are the light-shaded triangles. The resulting persistence diagrams are slightly better, especially for high values of the filtration parameter (that take into account the spurious light-shaded triangles). The truncated k-means succeed in removing a lot of centers accounting for noise (see Figure 2.5), but still involves centers far from K, and inside loop centers that blur the 1-persistence signal, resulting in a poor topological inference.
Next, we compute the centers and the corresponding distance to K estimator for the trimmed k-means and trimmed k-PDTM, with trim level α_0 and for $k = 50$. In this case, there is no need to remove code points with empty cell, since all cells are structurally non-empty.
Trimmed k-means succeed in removing far from K centers, compared to truncated

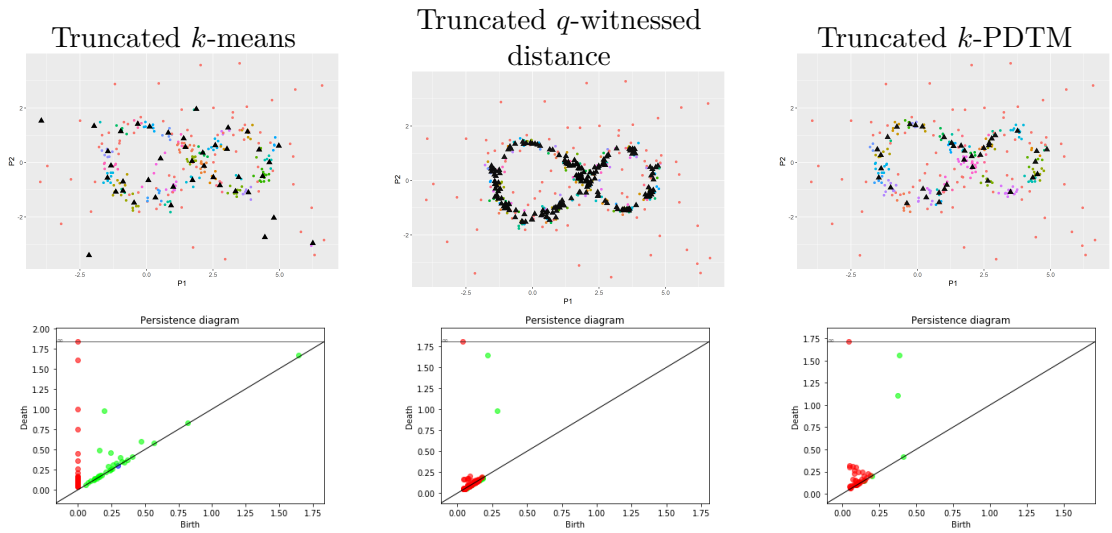


Figure 3.1: Comparison of truncation methods. Red points are labelled as outliers. In the persistence diagrams, red points correspond to connected components, green points to loops.

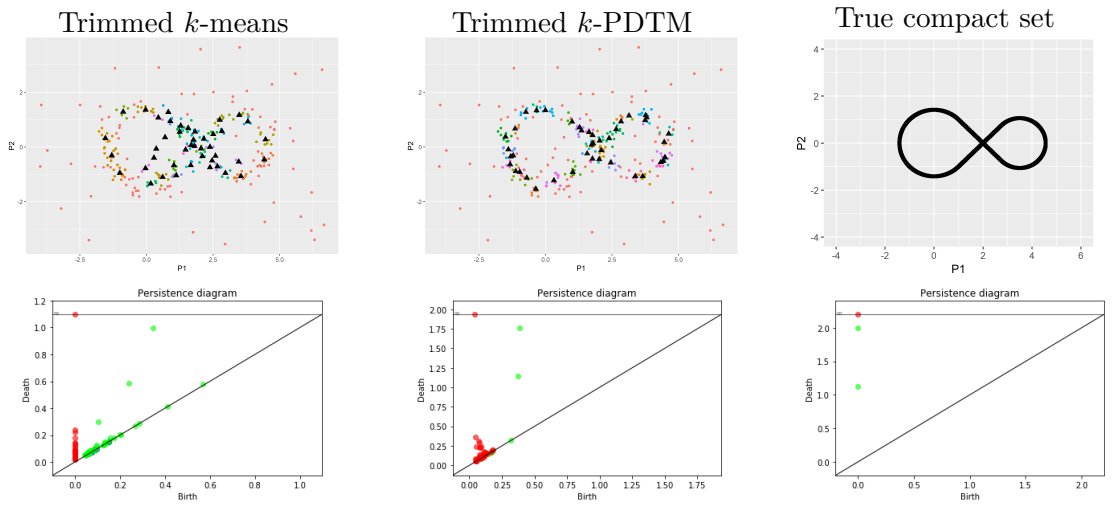
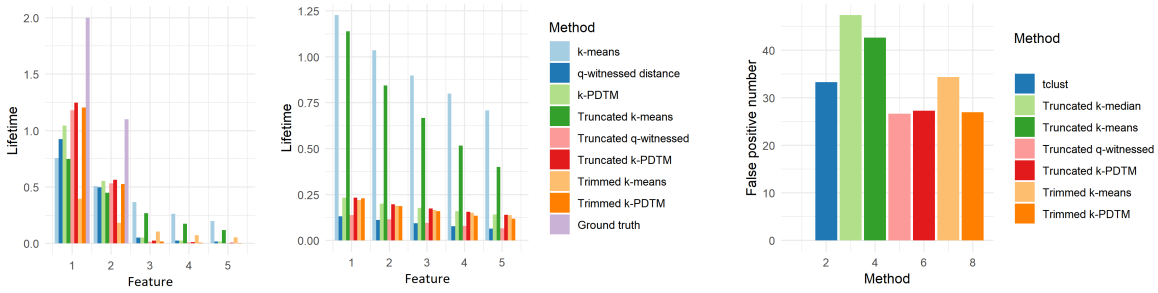


Figure 3.2: Trimmed k-PDTM and k-means. Red points are labelled as outliers.

k-means (see Figure 3.1). However, it also provides inside loop centers that do not allow to retrieve the two loops of K. Trimmed k-PDTM illustration is very close to the truncated k-PDTM one depicted in Figure 3.1. Difference between the two methods is very thin, and essentially consists in a slightly lesser persistence of spurious connected components for the trimmed k-PDTM, as exposed in Figure 3.3 below. It is interesting to note that $\alpha_0 = 2/7 >> 1/k = 0.02$, so that from a theoretical viewpoint the FSBP of the α-trimmed procedures cannot be controlled by Corollary 3.15. Nonetheless the outliers labellings these procedure provide remain relevant, showing that the prescribed bound on the outlier proportion given by the discernability factor is suited for the adversarial case, and might be too pessimistic for Huber's contamination model with noise less concentrated than signal.

To numerically assess the performance of our robust quantization procedures in outlier identification and topological inference, we repeated the experiment 100 times. At each time, we computed the lifetimes of the topological features and sorted them in decreasing order. Figure 3.3 (Left) below exposes the means of these lifetimes. Figure 3.3 (Right) depicts the mean amount of False positive over the 100 repetitions, that is

the number of signal points that are labeled as outliers by the algorithm. In addition to q-witness, k-means and k-PDTM, we also include comparison with other trimming approaches for outlier detection, such as tclust Fritz et al. [2012] (`tclust` function in `trimcluster R` library) and a truncated version of k-median Cardot et al. [2013] (`kGmedian` function of the `Gmedian R` library).



Holes Connected components False positive number

Figure 3.3: Features lifetimes and False positive numbers.

Overall, methods based on the DTM (q-witnessed, k-PDTM) outperform Euclidean-based methods (k-means, k-median, tclust). Concerning topological inference, k-means based methods show poor performance in retrieving 1-dimensional homology: no clear gap appears between the life spans of the two first features and the remaining spurious ones. For the 0-dimensional homology (where the first 0-dimensional feature, corresponding to the infinite connected component, has been removed in Figure 3.3), (truncated) k-means add too many persistent connected components compared to other methods. On the whole, k-PDTM-based approaches compared well with q-witnessed methods.
Concerning outlier detection, around 30 points drawn from the noise distribution fall close to K, so that all methods fail to detect them. q-witnessed distance, truncated and trimmed k-PDTM overall succeed in retrieving the remaining 50 noise points, whereas k-median and k-means based approaches correctly identify at least 10 points less than the aforementioned methods. tclust exhibits an intermediate behavior, correctly labeling around 45 outliers. For the q-witnessed and k-PDTM methods, labeling noise points that fall close to K as signal (hence removing some signal points that may be far from K) seems not to affect the topological correctness of the final distance estimator $\hat{\mathrm{d}}$. Such a behavior is likely to be due to the particular structure of the noise (Huber noise contamination with ambient noise). This might not be the case when encountering more structured noise, as in the following Section 3.1.6.

3.1.6 Robust Clustering

The robust Bregman quantization scheme introduced in the previous section finds applications in clustering. We expose here a particular application in stylometric clustering based on texts. The aim of clustering differs from the quantization one in the following sense: following Duda, Hart and Stork's formulation, given a set of points $\mathbb{X}_n$, *clustering* intends to partition $\mathbb{X}_n$ into dissimilar groups of similar items [Duda et al., 2001, Chapter 10]. Thus, clustering is focused on *cells*, while quantization also seeks for code points that represent each cell.
From a theoretical viewpoint, if we assume that $\mathbb{X}_n$ is drawn from a distribution Q, most existing clustering techniques are *density-based*, *centroid-based*, or *model-based*.
Density-based methods aim to partition data with respect to properties of the density q of Q (assuming that there exists a density), by choosing cells as super-level sets of the density (`DBSCAN` Ester et al. [1996], `Single Linkage` Gower and Ross [1969], `Spectral`

`Clustering` Ng et al. [2001]) or as basins of attraction of local modes (`MeanShift` Cheng [1995], `Tomato` Chazal et al. [2013]). Centroid-based methods are in fact based on the quantization principle: code cells with a prescribed shape and code points are built to minimize a distortion-like criterion. This class of methods encompasses Bregman clustering Banerjee et al. [2005a], but also non-Bregman dissimilarities such as L_1-norm (k-`median` MacQueen [1967], Cardot et al. [2012], k-`medoids` (Kaufman and Rousseeuw [1990]) and prescribed forms of cells such as rectangles (`CART`, Breiman et al. [1984]). At last, model-based clustering schemes intend to fit a mixture model distribution to data, then to partition the space via the MAP rule. The `EM` algorithm Dempster et al. [1977] is suited for the Gaussian mixture model. As detailed below, Bregman clustering algorithms have connections with model-based techniques, whenever the mixture components are assumed to be in an exponential family.

When some outliers are added, some of the aforementioned clustering methods can adapt to a certain extent. Namely, density-based clustering schemes are allegedly robust to some noise in Huber's contamination model, whereas some centroid-based methods such as k-`median` or k-`medoids` seem able to face some adversarial noise due to the L_1 criterion. Note that density-based methods are likely to show deprecated performances in high-dimensional settings, as for every non-parametric method. Apart from our robust Bregman quantization scheme, some other robustification methods for the k-`means` algorithm may be found in Klochkov et al. [2020], Brunet-Saumard et al. [2020], based on the Median of Means principle (see, e.g., Lecué and Lerasle [2020]). At last, we mention that the `tclust` algorithm Fritz et al. [2012] may be thought of as a robust version of the `EM` algorithm.

3.1.6.1 Bregman clustering and exponential mixtures

Recall that an exponential family associated to a proper closed convex function ψ defined on an open parameter space $\Theta \subset \mathbb{R}^d$ is a family of distributions $\mathcal{F}_\psi = \{P_{\psi,\theta} \mid \theta \in \Theta\}$, such that, for all $\theta \in \Theta$, $P_{\psi,\theta}$, defined on $\mathbb{R}^d$, is absolutely continuous with respect to some distribution P_0, with density $p_{\psi,\theta}$ defined for all $x \in \Omega$ by

$$p_{\psi,\theta}(x) = \exp(\langle x, \theta \rangle - \psi(\theta)).$$

The function ψ is called the cumulant function and θ is the natural parameter. For this model, the expectation of $P_{\psi,\theta}$ may be expressed as $\mu(\theta) = \nabla_\theta \psi$. We define

$$\phi(\mu) = \sup_{\theta \in \Theta} \{\langle \mu, \theta \rangle - \psi(\theta)\}.$$

By Legendre duality, for all μ such that ϕ is defined, we get $\phi(\mu) = \langle \theta(\mu), \mu \rangle - \psi(\theta(\mu))$, with $\theta(\mu) = \nabla_\mu \phi$. The density of $P_{\psi,\theta}$ with respect to P_0 can be rewritten using the Bregman divergence associated to ϕ as follows:

$$p_{\psi,\theta}(x) = \exp(-\mathrm{d}_\phi(x, \mu) + \phi(x)).$$

Standards distributions such as Gaussian, Poisson, Binomial and Gamma may be expressed that way. Table 3.1 below presents the 4 densities together with the functions ψ and ϕ, as well as the associated Bregman divergences d_ϕ.

Distribution	$p_{\psi,\theta}(x)$	θ	$\psi(\theta)$
Gaussian	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)$	$\frac{a}{\sigma^2}$	$\frac{\sigma^2}{2}\theta^2$
Poisson	$\frac{\lambda^x \exp(-\lambda)}{x!}$	$\log(\lambda)$	$\exp(\theta)$
Binomial	$\frac{N!}{x!(N-x)!}q^x(1-q)^{N-x}$	$\log\left(\frac{q}{1-q}\right)$	$N\log\left(1+\exp(\theta)\right)$
Gamma	$\frac{x^{k-1}\exp(-\frac{x}{b})}{\Gamma(k)b^k}$	$-\frac{k}{\mu}$	$k\log\left(-\frac{1}{\theta}\right)$

Distribution	μ	$\phi(\mu)$	$\mathrm{d}_\phi(x,\mu)$
Gaussian	a	$\frac{1}{2\sigma^2}\mu^2$	$\frac{1}{2\sigma^2}(x-\mu)^2$
Poisson	λ	$\mu\log(\mu)-\mu$	$x\log\left(\frac{x}{\mu}\right)-(x-\mu)$
Binomial	Nq	$\mu\log\left(\frac{\mu}{N}\right)+(N-\mu)\log\left(\frac{N-\mu}{N}\right)$	$x\log\left(\frac{x}{\mu}\right)+(N-x)\log\left(\frac{N-x}{N-\mu}\right)$
Gamma	kb	$-k+k\log\left(\frac{k}{\mu}\right)$	$\frac{k}{\mu}\left(\mu\log\left(\frac{\mu}{x}\right)+x-\mu\right)$

Table 3.1: Exponential family distributions and associated Bregman divergences.

More instances of exponential families and their related Bregman divergence can be found in [Fischer, 2010, Table 1]. It falls under the intuition that clustering with the appropriate Bregman divergence is relevant whenever the source distribution follows a mixture distribution with components that belong to the corresponding exponential family. In this case, clustering with Bregman divergence may be thought of as a hard-threshold model-based clustering scheme (see, e.g., Banerjee et al. [2005b]). The following Remark 3.17 gives an illustration of this connection in a simple case.

Remark 3.17. We let $k=2$, $\theta_1\neq\theta_2$, $z_1^*,\dots,z_n^*$ be hidden labels in $\{1,2\}$, and $X_1,\dots,X_n$ be an independent sample such that X_i has density

$$\mathbb{1}_{z_i^*=1}p_{\psi,\theta_1}(x)+\mathbb{1}_{z_i^*=2}p_{\psi,\theta_2}(x),$$

where $p_{\psi,\theta_j}(x)=\exp(-\mathrm{d}_\phi(x,\mu_j)+\phi(x))$, for $j\in\{1,2\}$. The parameters of this model are $(z_i^*)_{i\in[\![1,n]\!]},\theta_1,\theta_2$. This model slightly differs from a classical mixture model since the labels are not assumed to be drawn at random.
Let $z_{i,j}$, $i\in[\![1,n]\!]$, $j\in\{1,2\}$, denote assignment variables, that is such that $z_{i,j}=1$ if X_i is assigned to class j and 0 otherwise. Also denote by $m=\sum_{i=1}^n z_{i,1}$, $n-m=\sum_{i=1}^n z_{i,2}$, $\bar{X}_1=\sum_{i=1}^n X_i z_{i,1}/m$, $\bar{X}_2=\sum_{i=1}^n X_i z_{i,2}/(n-m)$. Maximizing the log-likelihood of the observations boils down to maximizing in $(z_{i,j})_{i,j}$:

$$\ln\prod_{i=1}^n\exp\left[-z_{i,1}\mathrm{d}_\phi\left(X_i,\bar{X}_1\right)-z_{i,2}\mathrm{d}_\phi\left(X_i,\bar{X}_2\right)+\phi(X_i)\right]$$
$$=-\sum_{i=1}^n z_{i,1}\mathrm{d}_\phi\left(X_i,\bar{X}_1\right)-\sum_{i=1}^n z_{i,2}\mathrm{d}_\phi\left(X_i,\bar{X}_2\right)+\sum_{i=1}^n\phi(X_i).$$

On the other hand, since optimal codebooks are local means of their Bregman-Voronoi cells (Proposition 3.7), minimizing $P_n\mathrm{d}_\phi(.,\mathbf{c})$ is equivalent to minimizing $\sum_{i=1}^n z_{i,1}\mathrm{d}_\phi\left(X_i,\bar{X}_1\right)+\sum_{i=1}^n z_{i,2}\mathrm{d}_\phi\left(X_i,\bar{X}_2\right)$. Thus, clustering with Bregman divergences is the same as maximum likelihood clustering based on this model. Further, if we assume that μ_1 and μ_2 are known, then the Bregman assignment rule $x\mapsto\arg\min_{j\in\{1,2\}}\mathrm{d}_\phi(x,\mu_j)$ is the Bayes rule.

3.1.6.2 A case study: stylometric author clustering

We illustrate the relevance of our robust Bregman quantization scheme for texts clustering based on stylometric descriptors, as described in [Taylor Arnold, 2015, Section 10]. Raw data consist in 26 annotated texts from 4 authors (Mark Twain, Sir Arthur

Conan Doyle, Nathaniel Hawthorne and Charles Dickens). These texts are available as supplementary material for Taylor Arnold [2015], and are framed as a sequence of lemmatized string characters (for instance "be" and "is" are instances of the same lemma "be"). Following Taylor Arnold [2015], we base our stylometric comparison on lemmas corresponding to nouns, verbs and adverbs, and split every original text in chunks of size 5000 of such lemmas that will be considered as data points. Then the 50 overall most frequent lemmas are chosen, and every chunk is described as the vector of counts of these lemmas within it. Thus, signal points consists of 189 count vectors with dimension 50, originating from 4 different authors.
The signal points are corrupted using the same process for the 8 State of the Union Addresses given by Barack Obama (available in `obama` dataset from package CleanNLP in R), resulting in 5 additional points, and for the King James Version of the Bible (available on Project Gutenberg) that we preliminary lemmatize using the CleanNLP package, resulting in 15 more additional points. Our final dataset consists of the $n = 189$ signal points and the $s = 20$ outlier points described above. Slightly anticipating, these 20 outliers might also be thought of as two additional small clusters with size 5 and 15. Since every individual lemma count can be modeled as a Poisson random variable in the random character sequence model Evert [2004], the appropriate Bregman divergence for this dataset is likely to be the Poisson divergence. In the following, we compare our method with Poisson divergence to trimmed k-means, trimmed k-medians, and t-clust.

Calibration of α and k:
When the number of clusters k is known beforehand, we propose the following heuristic to select the trimming parameter $q = (n + s)(1 - \alpha)$, that is, the number of points in the sample which are assigned to a cluster and not considered as noise. We let q vary from 1 to the sample size $n + s$, plot the curve $q \mapsto cost[q]$ where $cost[q]$ denotes the optimal empirical distortion at trimming level q, and choose $q^\star$ by seeking for a cut-point on the curve. Indeed, when the parameter q gets large enough, it is likely that the procedure begins to assign outliers to clusters, which dramatically deprecates the empirical distortion. Note that such an approach can also be used to calibrate α in a quantization framework: for k large enough, if the q kept points correspond to signal points from an (a, b)-standard distribution, then Zador's Theorem ([Graf and Luschgy, 2000, Theorem 6.2], [Liu and Belkin, 2016, Lemma2]) suggests that $cost[q] \sim qk^{-\frac{2}{b}}$, so that whenever some outliers from Huber's noise contamination model are added, the cost slope is likely to increase, to some extent depending on the difference between dimensionality of the signal and dimensionality of the noise (provided the noise has higher dimensionality than signal).
Whenever both k (number of clusters) and q are unknown, we propose to select these two parameters following the same principle as the algorithm `tclust` Fritz et al. [2012]. First we draw, for different values of k, the cost curves $q \mapsto cost_k[q]$, for $1 \leq q \leq n$. For each curve, the q's for which there is an abrupt slope increase can correspond to cases where outliers are assigned to clusters, or where some small clusters are included in the set of signal points (if k is chosen too small). In the sequel, we split $[\![1, n + s]\!]$ into several bins $[\![q_j, q_{j+1}]\!]$. On every such bin, we select a k that provides a significant cost decrease, as well as the q yielding a slope jump. Note that this heuristic may result in several possible pairs (k, q), corresponding to different point of views, depending on what data point are considered as outliers or not. An illustration of this fact is given in Figure 3.4, where outliers consist in small additional clusters.
In Figure 3.4, we draw the cost of our method as a function of q (number of kept points, that is $(n + s)(1 - \alpha)$), for different cluster numbers k. According to this figure, several choices of k and q are possible. For values of q up to 175, the significant jumps in the risk function are for $k = 3$ and $k = 6$. For $k = 3$, the slope heuristic yields $q = 175$, whereas for $k = 6$ the slope heuristic suggests that no data points might be considered as

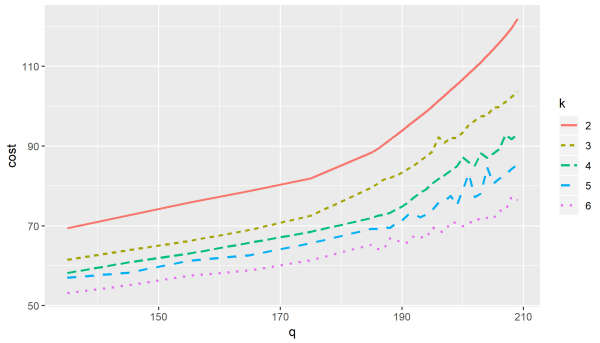


Figure 3.4: Cost curves for authors clustering with Poisson divergence, with respect to number of kept points q.

outliers. When q ranges between 175 and 193, the significant distortion jumps are for $k = 4$ and $k = 6$, another possible choice is then $k = 4$ and $q = 188$. When q is larger than 193, the only significant jump is for $k = 6$. To summarize, the pairs $(k = 3, q = 175)$, $(k = 4, q = 188)$, $(k = 6, q = n = 209)$ seem reasonable. These three solutions correspond to the 3 natural trimmed partitions: clustering only 3 authors writings (Twain writings being considered as outliers), clustering the 4 authors writings and removing the outliers from the Bible and B. Obama addresses, and at last clustering the six sources of writings (none of them being considered as noise). The two latter situations are depicted in Figure 3.5, in the 2-dimensional basis given by a linear discriminant analysis of the proposed clustering.

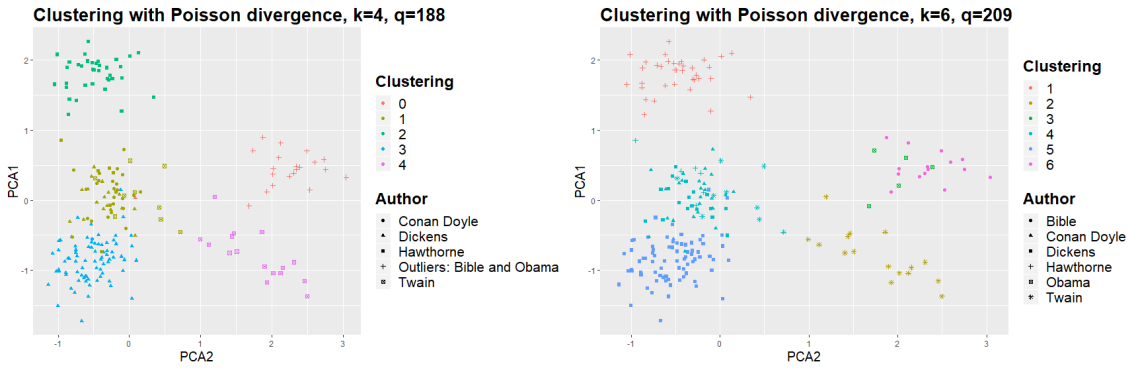


Figure 3.5: Author stylometric clustering with Poisson divergence.

Quality of clustering schemes will be assessed in terms of their Normalized Mutual Information with respect to ground truth (NMI, Strehl and Ghosh [2002]), that ranges between 0 ("orthogonal" clustering to ground truth) and 1 (same partition as ground truth). The outliers are treated as the 0-th cluster. For $k = 6$ and $q = 209$, our clustering globally retrieves the corresponding author. When $k = 4, q = 188$ is chosen, outliers are correctly identified and only one sample text from C. Dickens is labeled as outlier. The sample points seems on the whole well classified, that is assessed by a NMI of 0.7347. This performance is compared with the other clustering algorithms in Table 3.2. Note that values of q for competitors have been chosen to minimize the NMI, leading to $q = 190$ for trimmed k-means, $q = 202$ for trimmed k-medians, and $q = 184$ for `tclust`. The corresponding NMI curves may be found in [Brécheteau et al., 2021, Supplementary material].

The associated partitions for k-`median` and `tclust` are depicted in Figure 3.6, showing that these two methods fail in correctly identifying outliers.

Method	trimmed 4-means	trimmed 4-medians	tclust	trimmed Poisson
NMI	0.5336	0.4334	0.4913	0.7347

Table 3.2: Comparison of robust clustering methods for Author retrieving.

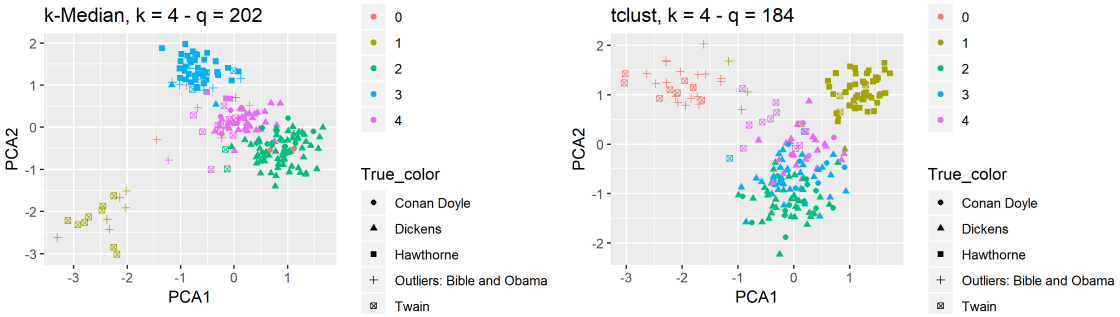


Figure 3.6: Author stylometric clustering with trimmed k-median and tclust.

On the whole, the Bregman clustering scheme with Poisson divergence seems well suited to the generative model of stylometric descriptors described in Evert [2004]. The trimmed version we propose allows to adapt to structured noise (such as spurious clusters). It also provides a screening method to investigate what may be considered as noise (via the calibration of k and α), at the cost of more expensive computations.

3.2 Clustering of measures

Let us conclude this chapter with an other application of the quantization principle to clustering. We intend here to cluster measures, that is, given a sample $\mathbb{X}_n = \{X_1, \dots, X_n\}$ where X_j is a (finite) measure on $\mathbb{R}^D$, to provide a partition of $\mathbb{X}_n$ into L classes. The proposed procedure relies on a vectorization of the measures, that is a map from $\mathcal{M}(\mathbb{R}^D)$ onto $\mathbb{R}^k$ build via quantization that preserves cluster structures if any. From a practical viewpoint, clustering of measures finds a natural application in clustering persistence diagrams (each multiset of points in $\mathbb{R}^2$ described in Section 2.1.1.1 may be thought of as a finite measure), but also in every situation where data comes as a collection of finite point sets, for instance in ecology Renner et al. [2015], genetics Royer et al. [2021], Adams et al. [2017], graphs clustering Carrière et al. [2019], Hoan Tran et al. [2018] and shapes clustering Chazal et al. [2009].
The vectorization scheme we propose is simple: for a codebook $\mathbf{c} \in (\mathbb{R}^D)^k$, we map each measure X_i onto a vector $v_i \in \mathbb{R}^k$ that roughly encodes how much mass X_i spreads around each code point. We build such a codebook via the quantization of a central tendency of the X_i's, the so-called mean measure, in Section 3.2.1. Then we show in Section 3.2.2 that for mixture models and under suitable assumptions such a vectorization allows to discriminate between mixture components.

3.2.1 Quantization of the mean measure

Assuming that the sample measures X_i's are i.i.d. with distribution X, the targeted central tendency is the mean measure $\mathbb{E}(X)$, defined by $(\mathbb{E}(X))(A) = \mathbb{E}(X(A))$, for any measurable set $A \subset \mathbb{R}^D$. We intend to approximate $\mathbb{E}(X)$ by a k-points supported measure via quantization, that is minimizing in $\mathbf{c}$ a distortion-like criterion

$$R(\mathbf{c}) = \mathbb{E}(X)(du) \min_{j=1,\dots,k} \|u - c_j\|^2 = W_2^2(\mathbb{E}(X), P_{\mathbf{c}}), \tag{3.4}$$

where $P_{\mathbf{c}} = \sum_{j=1}^{k} \mathbb{E}(X)(C_j(\mathbf{c}))\delta_{c_j}$, whenever $\mathbb{E}(X)(du)\|u\|^2 < +\infty$. Note that, if $P'_{\mathbf{c}} = \sum_{j=1}^{k} \mu_j \delta_{c_j}$, with $\sum_{j=1}^{k} \mu_j = \mathbb{E}(X)(\mathbb{R}^D)$, then $W_2(\mathbb{E}(X), P'_{\mathbf{c}}) \geq W_2(\mathbb{E}(X), P_{\mathbf{c}})$, so that finding an optimal codebook in the sense of (3.4) amounts to find an optimal k-points approximation of $\mathbb{E}(X)$ in terms of Wasserstein distance.
From now on we restrict our attention to bounded support and bounded total mass measures. Denoting by $\mathcal{M}(\Lambda, M)$ the set of measures Q such that $Q(\mathbb{R}^D) \leq M$ and $\mathrm{Supp}(Q) \subset \mathrm{B}(0, \Lambda)$, we assume that $X \in \mathcal{M}(\Lambda, M)$ a.s., so that $\mathbb{E}(X) \in \mathcal{M}(\Lambda, M)$. In particular, the distortion R of (3.4) is well-defined, and [Fischer, 2010, Corollary 3.1] ensures that there exist optimal codebooks in the sense of (3.4), whose set will be denoted by $\mathcal{C}_{opt}$.
Based on $\mathbb{X}_n$, a straightforward approach to build from sample a k-points approximation is to minimize $\mathbf{c} \mapsto \bar{X}_n(du) \min_{j=1,\dots,k} \|u - c_j\|^2$, where $\bar{X}_n := \frac{1}{n} X_i$ (since $X \in \mathcal{M}(\Lambda, M)$ a.s., $\bar{X}_n \in \mathcal{M}(\Lambda, M)$ a.s.). Since minimizing such a criterion is NP-hard in general (see, e.g., Mahajan et al. [2012]), we propose a stochastic-gradient like algorithm that we prove (almost) minimax optimal, under suitable assumptions.
Note that other central tendency measures and k-points approximations are possible. For instance, one may choose as central measure a restrained Wasserstein barycenter of the form $\arg\min_{\nu, |\mathrm{Supp}(\nu)| \leq k} \mathbb{E}\left(W_2^2(X, \nu)\right)$ and its empirical counterpart. However, such an approach is valid only in the case where $X(\mathbb{R}^D)$ is constant. Moreover, computing such a Wasserstein barycenter for $X_1, \dots, X_n$ becomes intractable for large n's, even with approximating algorithm Cuturi and Doucet [2013], Rabin et al. [2012]. The mean measure $\mathbb{E}(X)$ and its empirical counterpart are much more easier to compute, and can be of particular interest in image analysis Cuturi and Doucet [2013] or point processes modeling Renner et al. [2015], Diggle [1990], Shirota et al. [2017] (in this case the mean measure is referred to as the *intensity function* of the point process).
To minimize the empirical distortion, an adaptation of Lloyd's Algorithm Lloyd [1982] or Algorithm 1 is possible (see [Chazal et al., 2021, Algorithm 1]). We rather expose here a counterpart of Mac Queen's algorithm MacQueen [1967], based on mini-batches. Recall that $\pi_{\mathrm{B}(0,\Lambda)}$ denote the projection onto $\mathrm{B}(0, \Lambda)$.

Algorithm 2. Mini-batch k-means

- **Input:** $\{X_1, \dots, X_n\} = \mathbb{X}_n$, divided into mini-batches $(B_1, \dots, B_T)$ with sizes $(n_1, \dots, n_T)$, and k.
  For $t = 1, \dots, T$, B_t is divided in two halves , $B_t^{(1)}$ and $B_t^{(2)}$;
- **Initialization:** Sample $c_1^{(0)}$, $c_2^{(0)}$,... $c_k^{(0)}$ from $\bar{X}_n$.
- **Iterations:** **for** $t = 0, \dots, T-1$ do:
  - **for** $j = 1, \dots, k$,

$$c_j^{(t+1)} = \pi_{\mathrm{B}(0,\Lambda)} \left(c_j^{(t)} - \frac{\bar{X}_{B_{t+1}^{(2)}}(du) \left[(c_j^{(t)} - u) \mathbb{1}_{C_j(\mathbf{c}^{(t)})}(u) \right]}{(t+1) \bar{X}_{B_{t+1}^{(1)}}\left(C_j(\mathbf{c}^{(t)})\right)} \right);$$

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

We mention here that the mini-batches split in two halves is motivated by technical considerations only. In the following experiments we take $B_t = B_t^{(1)} = B_t^{(2)}$. Whenever $X_i = \delta_{x_i}$ for $i = 1, \dots, n$ (and $B_i^{(1)} = B_i^{(2)} = \{x_i\}$), that is the classical point sample case, Algorithm 2 is a slight modification of the original Mac-Queen algorithm MacQueen [1967]. Indeed, the Mac-Queen algorithm takes mini-batches of size 1, and estimates the population of the cell j at the t-th iteration via $\sum_{\ell=1}^{t} \hat{p}_j^{(\ell)}$ instead of $t\hat{p}_j^{(t)}$,

where $\hat{p}_j^{(t)} = \bar{X}_{B_t^{(1)}}\left(C_j(\mathbf{c}^{(t)})\right)$. These modifications are motivated by Theorem 3.21, that guarantees near-optimality of the output of Algorithm 2, provided that the mini-batches are large enough and that $\mathbb{E}(X)$ satisfies the assumption described below.

To formally introduce the so-called margin condition, additional quantities must be introduced. In what follows, R^* will denote the optimal distortion achievable with k points, that is $R^* = R(\mathbf{c})$, where $\mathbf{c} \in \mathcal{C}_{opt}$. It is immediate that $R^* = 0$ if and only if $\mathbb{E}(X)$ is supported by less than k points, in which case the quantization problem is trivial. In what follows we assume that $\mathbb{E}(X)$ is supported by more than k points. Some basic properties of $\mathcal{C}_{opt}$ are needed.

PROPOSITION 3.18 : [LEVRARD, 2018, PROPOSITION 1]

Recall that $\mathbb{E}(X) \in \mathcal{M}(\Lambda, M)$. *Then,*

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

We can now introduce a so-called *margin condition*, that slightly extends the one exposed in [Levrard, 2015, Definition 2.1] for the point sample case.

DEFINITION 3.19 : MARGIN CONDITION

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

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

where, for $\mathbf{c} \in (\mathbb{R}^D)^k$,

$$N(\mathbf{c}) = \bigcap_{i \neq j} V_i(\mathbf{c}) \cap V_j(\mathbf{c}),$$

that is the skeleton of the Voronoi diagram of $\mathbf{c}$.

Intuitively speaking, a margin condition ensures that the mean distribution $\mathbb{E}(X)$ is well-concentrated around k poles. For instance, finitely-supported distributions satisfy a margin condition. Following Levrard [2018], a margin condition will ensure that usual k-means type algorithms are almost optimal in terms of distortion. Up to our knowledge, margin-like conditions are always required to guarantee convergence of Lloyd-type algorithms Tang and Monteleoni [2016], Levrard [2018]. From a technical viewpoint, the margin condition ensures that the the risk function defined by (3.4) is locally co-coercitive around optimal codebooks.

LEMMA 3.20 : [CHAZAL ET AL., 2021, LEMMA 21]

Assume that $\mathbb{E}(X)$ satisfies a margin condition with radius r_0. Denoting by $\Lambda_0 = \frac{Br_0}{16\sqrt{2}\Lambda}$, if $\mathbf{c}^* \in \mathcal{C}_{opt}$ and $\|\mathbf{c} - \mathbf{c}^*\| \leq \Lambda_0$, it holds

$$\left\langle D(\mathbf{c})^{-1} G(\mathbf{c}), \mathbf{c}^* - \mathbf{c} \right\rangle \geq \left(\frac{64}{65} - \frac{1}{8\sqrt{2}} \right) \|\mathbf{c} - \mathbf{c}^*\|^2,$$

where $G(\mathbf{c})$ is the gradient of R at $\mathbf{c}$, defined by

$$G(\mathbf{c})_j = 2\mathbb{E}(X)(du)[(u - c_j)\mathbb{1}_{C_j(\mathbf{c})}(u)],$$

and $D(\mathbf{c})$ is the diagonal matrix with entries $2\mathbb{E}(X)\left(C_j(\mathbf{c})\right)$, for $j = 1, \ldots, k$, provided that $\mathbb{E}(X)\left(C_j(\mathbf{c})\right) \neq 0$.

Of course the constants $64/65$, $1/8\sqrt{2}$ are purely artificial, and depend on the radius Λ_0. The main idea is that locally, around every optimal codebook $\mathbf{c}^*$, it holds $\langle G(\mathbf{c}), \mathbf{c}^* - \mathbf{c} \rangle \geq a\|\mathbf{c} - \mathbf{c}^*\|^2$, for a fixed constant a. In particular, this ensures that codebooks in $\mathcal{C}_{opt}$ are Λ_0-separated. Such a co-coercitivity condition is a key ingredient to prove convergence of stochastic-gradient based techniques to approximate a minimizer of R. Such techniques apply to Algorithm 2, that may be considered as a mini-batch stochastic gradient descent.

THEOREM 3.21 : [CHAZAL ET AL., 2021, THEOREM 10]

Let $X \in \mathcal{M}(\Lambda, M)$. Assume that $\mathbb{E}(X)$ satisfies a margin condition with radius r_0, and denote by $\Lambda_0 = \frac{Br_0}{16\sqrt{2}\Lambda}$, $\kappa_0 = \Lambda_0/\Lambda$. If $(B_1, \ldots, B_T)$ are equally sized mini-batches of length $ckM^2 \log(n)/(\kappa_0 p_{\min})^2$, where c is a positive constant, and $\mathbf{c}^{(T)}$ denotes the output of Algorithm 2, then, provided that $\mathbf{c}^{(0)} \in \mathrm{B}(\mathcal{C}_{opt}, \Lambda_0)$, we have

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

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

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

that is without splitting the batches.

Interestingly, the result of Theorem 3.21 is dimension-free. In other words, this result would also be valid in any separable Hilbert space, provided that $\mathrm{Supp}(X) \subset \mathrm{B}(0, \Lambda)$ a.s., as for the point sample result [Levrard, 2015, Theorem 3.1]. The sample size dependency is optimal up to a $\log(n)$ factor: indeed, [Levrard, 2018, Proposition 7] provides a $1/n$ lower bound in the point sample case under margin condition, that is a sub-case of the measure sample (with margin condition). Note that the $1/n$ rate may be attained in the measure sample case using a Lloyd type algorithm (see, e.g., [Chazal et al., 2021, Theorem 9]), in the finite-dimensional case with margin condition. In this case, deviation bounds might also be stated, that are not easy to derive for Algorithm 2. The dependency on k is more subtle to investigate: in the point sample case, under margin condition with fixed parameters B, p_{min} and r_0, [Levrard, 2018, Proposition 7] provides a lower bound in $k^{1-2/D}$, whereas [Levrard, 2015, Theorem 3.1] states an upper

bound in k. From this viewpoint the k^2 term of Theorem 3.21 appears suboptimal. However, it is important to remark that quantities such as $p_{\min}$ and κ_0 also depend on k, so that the overall influence of k remains a hard question to answer.
As a byproduct, Theorem 3.21 shed some light on what a good initialization is. Roughly speaking, if the initial state of Algorithm 2 is in the area where the distortion is convex, then the output has the desired behavior. The size of such a good initialization neighborhood depends on the margin condition, and a lower bound on the volume of $\mathrm{B}(\mathcal{C}_{opt}, \Lambda_0)$ may be used to prescribe the number of trials needed to get at least one good initialization in practice. This intuition may be compared with the good initializations for Lloyd's algorithm given in Tang and Monteleoni [2016], that need to be close to an optimal from a distortion point of view. However such a volume approach will heavily suffer from the curse of dimensionality, the volume of $\mathrm{B}(\mathbf{c}^*, \Lambda_0)$ being of order $(\Lambda_0/\Lambda)^D$ whenever $\mathbb{E}(X)$ has a D-dimensional density. In the following experiments, several initializations with a k-`means ++`-like Arthur and Vassilvitskii [2007] approach are drawn, and the final codebook with the lowest empirical distortion is chosen.
At last, though natural in a clustering framework, the margin condition exposed in Definition 3.19 seems not necessary for Lloyd and MacQueen algorithm to converge. Indeed, imposing structural assumptions such as log-concavity of a density is enough to provide a co-coercitivity result such as Lemma 3.20, hence to guarantee convergence of the two aforementioned algorithms in dimension 1 (Kieffer [1983]). Extending the validity of results such as Theorem 3.21 is part of the ongoing work of the authors. A preliminary result shows that the log-concavity assumption of Kieffer [1983] can be removed, proving that the optimal $1/n$ convergence rate may be attained generally in dimension $D = 1$, whenever $\mathbb{E}(X)$ has a density. The authors strongly believe that such a result can be extended to higher dimensions.
Once the mean measure being (almost optimally) quantized, we can now use this k-points approximation to build a measure vectorization scheme.

3.2.2 Vectorization of measures and clustering

For a given codebook $\mathbf{c}$ and a scale $r > 0$, we may represent a measure μ via the vector of weights $(\mu(\mathrm{B}(c_1, r)), \dots, \mu(\mathrm{B}(c_k, r)))$ that encodes the mass that μ spreads around every pole c_j. Provided that the codepoints are discriminative (this will be discussed in the following section), separation between clusters of measure will be preserved. In practice, convolution with kernels is often preferred to local masses (see, e.g., Royer et al. [2021]). To ease computation, we will restrict ourselves to the following class of kernel functions.

Definition 3.22 :

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

$$\begin{array}{llll} i) & \|\psi\|_\infty \leq 1, & ii) & \sup_{|u|\leq 1/p} \psi(u) \geq 1 - \delta, \\ iii) & \sup_{|u|>2p} \psi(u) \leq \delta, & iv) & \psi \text{ is 1-Lipschitz.} \end{array}$$

Let us mention that a (p, δ)-kernel is also a (q, δ)-kernel, for $q > p$. This definition of a kernel function encompasses widely used kernels, such as Gaussian or Laplace kernels. In particular, the function $\psi(u) = exp(-u)$ that is used in Royer et al. [2021] is a $(p, 1/p)$-kernel for $p \in \mathbb{N}^*$. The 1-Lispchitz requirement is not necessary to prove that the representations of two separated measures will be well-separated. However, it is a key assumption to prove that the representations of two measures from the same cluster will remain close in $\mathbb{R}^k$. From a theoretical viewpoint, a convenient kernel is $\psi_0 : x \mapsto (1 - ((x - 1) \vee 0)) \vee 0$, which is a $(1, 0)$-kernel, thus a $(p, 0)$-kernel for all $p \in \mathbb{N}^*$.

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

$$v_{\mathbf{c},\sigma} : \begin{cases} \mathcal{M}(\Lambda, M) & \to & [0,\Lambda]^k \\ \mu & \mapsto & (\mu(du)\psi(\|u-c_1\|/\sigma), \ldots, \mu(du)\psi(\|u-c_k\|/\sigma)) \,. \end{cases} \tag{3.5}$$

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

3.2.2.1 Discriminative codebooks

To assess that our vectorization scheme is discriminative in some sense, we have to assume that there exists hidden classes Z_i's and seek to retrieve them. Thus, from now on we assume that there exists a vector $(Z_1, \ldots, Z_n) \in [\![1, L]\!]^n$ of (hidden) label variables, and denote by $X^{(\ell)} \sim X \mid Z = \ell$, the distributions of the mixture components. We further assume that $X^{(\ell)} \in \mathcal{M}(\Lambda, M)$, for every $\ell \in [\![1, L]\!]$.
For a vectorization v, a good discrimination property should be to guarantee that $\|v(X_{i_1}) - v(X_{i_2})\|$ is large whenever $Z_{i_1} \neq Z_{i_2}$. For the vectorization $v_{\mathbf{c},\sigma}$ proposed by (3.5), discrimination between two classes $\mu_1 \neq \mu_2$ will likely occur whenever $\mu_1(\mathrm{B}(c_j, \sigma))$ is different enough from $\mu_2(\mathrm{B}(c_j, \sigma))$, for some c_j. Since there exists $x \in \mathrm{B}(0, \Lambda)$ and $r > 0$ such that $\mu_1(\mathrm{B}(x, r)) \neq \mu_2(\mathrm{B}(x, r))$, such a discrimination seems possible if there exists a codepoint c_j close enough to x and σ is of order r. We extend this intuition to the multi-classes case, by introducing, for a given codebook $\mathbf{c}$, the following definition of (p, r, Δ)-scattering to quantify how well $\mathbf{c}$ will allow to separate clusters.

Definition 3.23 :

> Let $(p, r, \Delta) \in (\mathbb{N}^* \times \mathbb{R}^+ \times \mathbb{R}^+)$. A codebook $\mathbf{c} \in \mathrm{B}(0, \Lambda)^k$ is said to (p, r, Δ) -**shatter** $X_1, \ldots, X_n$ if, for any $i_1, i_2 \in [\![1, n]\!]$ such that $Z_{i_1} \neq Z_{i_2}$, there exists $j_{i_1,i_2} \in [\![1, k]\!]$ such that
>
> $$X_{i_1}(\mathrm{B}(c_{j_{i_1,i_2}}, r/p)) \geq X_{i_2}(\mathrm{B}(c_{j_{i_1,i_2}}, 4pr)) + \Delta, \quad \text{or}$$
> $$X_{i_2}(\mathrm{B}(c_{j_{i_1,i_2}}, r/p)) \geq X_{i_1}(\mathrm{B}(c_{j_{i_1,i_2}}, 4pr)) + \Delta.$$

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

Proposition 3.24 :

> *Assume that* $\mathbf{c} \in \mathrm{B}(0, \Lambda)^k$ *shatters* $X_1, \ldots, X_n$*, with parameters* (p, r, Δ)*. Then, if* Ψ *is a* (p, δ)*-kernel, with* $\delta \leq \frac{\Delta}{4M}$*, we have, for all* i_1*,* $i_2 \in [\![1, n]\!]$ *and* $\sigma \in [r, 2r]$*,*
>
> $$Z_{i_1} \neq Z_{i_2} \quad \Rightarrow \quad \|v_{\mathbf{c},\sigma}(X_{i_1}) - v_{\mathbf{c},\sigma}(X_{i_2})\|_\infty \geq \frac{\Delta}{2}.$$

This proposition gives details on how $X_1, \ldots, X_n$ has to be shattered with respect to the parameters of Ψ. Indeed, assume that $\Delta = 1$ (that is the case if the X_i's are integer-valued measures, such as count processes for instance). Then, to separate

clusters, one has to choose δ small enough compared to $1/M$, and thus p large enough if Ψ is non-increasing. Hence, the vectorization will work roughly if the support points of two different counting processes are rp-separated, for some scale r. This scale r then drives the choice of the bandwidth σ. Section 3.2.3 below provides an instance of shattered measures in the case where measures are persistence diagrams originating from well separated shapes. If the requirements of Proposition 3.24 are fulfilled, then a standard hierarchical clustering procedure such as Single Linkage with L_∞ distance will separate the clusters for the scales smaller than $\Delta/2$. Without more assumptions, this global clustering scheme can result in more than L clusters, nested into the ground truth label classes.
Now, to achieve a perfect clustering of the sample based on the vectorization $v_{\mathbf{c},\sigma}$, we have to ensure that measures from the same cluster are not too far in terms of Wasserstein distance, implying in particular that they have the same total mass. This motivates the following definition.

Definition 3.25 :

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

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

where W_1 denotes the 1-Wasserstein distance.

It now falls under the intuition that well-concentrated and shattered sample measures are likely to be represented in $\mathbb{R}^k$ by well-clusterable points. A precise statement is given by the following Proposition 3.26.

Proposition 3.26 :

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

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

Therefore, if $X_1, \dots, X_n$ is (p, r, Δ)-shattered by $\mathbf{c}$, and $(r\Delta/4)$-concentrated, then, for any (p, δ)-kernel satisfying $\delta \leq \frac{\Delta}{4M}$, we have, for $\sigma \in [r, 2r]$,

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

An immediate consequence of Proposition 3.26 is that (p, r, Δ)-shattered and $r\Delta/4$-concentrated sample measures can be vectorized in $\mathbb{R}^k$ into a point cloud that is structured in L clusters. These clusters can be exactly recovered via Single Linkage clustering, with stopping parameter in $]\Delta/4, \Delta/2]$. In practice, tuning the parameter σ is crucial. Some heuristic is proposed in Royer et al. [2021] in the special case of i.i.d persistence diagrams. An alternative calibration strategy is proposed in Section 3.2.3. It remains to prove that optimal codebooks obtained in Section 3.2.1 are shattering codebooks. Intuitively speaking, for large k's it is likely that optimal codebooks have code points in discriminative areas. Providing upper bounds on such a suitable k is a case-dependent issue. An example is given in the following Section 3.2.3, in a mixture of persistence diagram framework.

3.2.3 Application to clustering of persistence diagrams

In this section we consider a sample $\mathbb{X}_n$ of n persistence diagrams (see Section 2.1.1.1), that we consider as a sample of finite measures on $H^+ = \{(x, y) \in \mathbb{R}^2 \mid y > x\}$ (see (3.6) below). For more insights on the measure point of view on persistence diagrams the interested reader is referred to Chazal and Divol [2019]. First we provide theoretical guarantees in the case where these diagrams follow a mixture distribution.

3.2.3.1 Clustering persistence diagrams of a mixture of shapes

Our mixture of persistence diagrams model is the following. For $\ell \in [\![1, L]\!]$ let $S^{(\ell)}$ denote a compact d_ℓ-manifold and $D^{(\ell)}$ denote the persistence diagram generated by $\mathrm{d}_{S^{(\ell)}}$, the distance to $S^{(\ell)}$. For a fixed scale $s > 0$, we define $D^{(\ell)}_{\geq s}$ as the thresholded persistence diagram of $\mathrm{d}_{S^{(\ell)}}$, that is

$$D^{(\ell)}_{\geq s} = \sum_{\{(b,d)\in D^{(\ell)} \mid d-b\geq s\}} n(b, d)\delta_{(b,d)} := \sum_{j=1}^{k_0^{(\ell)}} n(m_j^{(\ell)})\delta_{m_j^{(\ell)}}, \tag{3.6}$$

where the multiplicities $n(b, d) \in \mathbb{N}^*$, and the $m_j^{(\ell)}$'s satisfy $(m_j^{(\ell)})_2 - (m_j^{(\ell)})_1 \geq s$. Let us mention here that $k_0^{(\ell)}$ is finite, according to [Chazal et al., 2021, Lemma 12].
Now, for $\ell \in [\![1, L]\!]$ we let $\mathbb{Y}_{N_\ell}$ be an N_ℓ-sample drawn on $S^{(\ell)}$ with density $f^{(\ell)}$ satisfying $f^{(\ell)}(u) \geq f_{min,\ell}$, for $u \in S^{(\ell)}$. If $\hat{D}^{(\ell)}$ denotes the persistence diagram generated by $\mathrm{d}_{\mathbb{Y}_{N_\ell}}$, we define $X^{(\ell)}$ as a thresholded version of $\hat{D}^{(\ell)}$, that is

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

for some bandwidth h_ℓ. At last, recall that our mixture model is defined by $X_i \mid \{Z_i = \ell\} \sim X^{(\ell)}$, so that conditionally on $Z_i = \ell$, X_i is a thresholded persistence diagram corresponding to a N_ℓ-sampling of the shape $S^{(\ell)}$. Recall also that the weights of the components are defined by $\pi_\ell = \mathbb{P}(Z = \ell)$.
To retrieve the labels Z_i's from a vectorization of the X_i's, we have to assume that the persistence diagrams of the underlying shapes differ by at least one point.

Definition 3.27 :

> The shapes $S^{(1)}, \ldots, S^{(\ell)}$ are **discriminable at scale s** if for any $1 \leq \ell_1 < \ell_2 \leq L$ there exists $m_{\ell_1,\ell_2} \in H^+$ such that
>
> $$D^{(\ell_1)}_{\geq s}(\{m_{\ell_1,\ell_2}\}) \neq D^{(\ell_2)}_{\geq s}(\{m_{\ell_1,\ell_2}\}),$$
>
> where the thresholded persistence diagrams are considered as measures.

It is immediate that, if m_{ℓ_1,ℓ_2} satisfies the discrimination condition stated above, then $m_{\ell_1,\ell_2} \in D^{(\ell_1)}_{\geq s}$ or $m_{\ell_1,\ell_2} \in D^{(\ell_2)}_{\geq s}$. Next, we must ensure that optimal codebooks of the mean measure have codepoints close enough to discrimination points.

PROPOSITION 3.28 :

Let $h_\ell = \left(\frac{C_{d_\ell}\log(N_\ell)}{f_{min,\ell}N_\ell}\right)^{1/d_\ell}$, *for some constants* C_{d_ℓ}, *and* $h = \max_{\ell\le L} h_\ell$. *Moreover, let* $M_\ell = D^{(\ell)}_{\ge s}(H^+)$, $\bar{M} = \sum_{\ell=1}^L \pi_\ell M_\ell$, *and* $\pi_{min} = \min_{\ell\le L}\pi_\ell$.

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

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

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

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

The bandwidths h_ℓ are defined in line with the convergence rate for the point sample in the (a,d_ℓ)-standard case (see Theorem 1.4 and Proposition 1.16, Item (iii)). If $\bar{D}_{\ge s}$ denotes the mean persistence diagram $\sum_{\ell=1}^L \pi_\ell D^{(\ell)}_{\ge s}$, and $\bar{D}_{\ge s}$ has K_0 points, then it is immediate that $k_0 + K_0(h) \le K_0$. Moreover, we also have $k_0 \le \frac{L(L+1)}{2}$. At last, if the $S^{(\ell)}$'s are included in $\mathrm{B}(0,\Lambda)$, a standard covering argument entails $K_0(h) \le \frac{1}{2}\left(\frac{\Lambda}{h}\right)^2$. To sum up, the choice $k = K_0$ satisfies the requirements of Proposition 3.28, but smaller choices are possible depending on the structure of the mixture.
Proposition 3.28 ensures that the discrimination points are well enough approximated by optimal k-centers of the expected persistence diagram $\mathbb{E}(X)$, provided the shapes $S^{(\ell)}$ are well-enough sampled and k is large enough so that $\bar{D}_{\ge s}$ is well-covered by k balls with radius h. In turn, provided that the shapes $S^{(1)},\dots,S^{(L)}$ are discriminable at scale s and that k is large enough, we can prove that an optimal k-points codebook $\mathbf{c}^*$ is a (p,r,Δ)-shattering of the sample, with high probability.

PROPOSITION 3.29 :

Assume that the requirements of Proposition 3.28 are satisfied. Let $\tilde{B} = \min_{i=1,\dots,k_0, j=1,\dots,K_0, j\ne i}\|m_i - m_j\|_\infty \wedge s$. *Let* $\kappa > 0$ *be a small enough constant. Then, if* N_ℓ *is large enough for all* $\ell\in[\![1,\ell]\!]$, $X_1,\dots,X_n$ *is* $(p,r,1)$*-shattered by* $\mathbf{c}^*$, *with probability larger than* $1 - n\max_{\ell\le L} N_\ell^{-\left(\frac{(\kappa\tilde{B})^{d_\ell} f_{min,\ell}N_\ell}{C_\ell d_\ell \log(N_\ell)}\right)}$, *provided that*

- $\frac{r}{p} \ge 2\kappa\tilde{B}$,
- $4rp \le \left(\frac{1}{2} - \kappa\right)\tilde{B}$.

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

Proposition 3.29 provides scales for the shattering constants of $\mathbf{c}^*$ in terms of the mixture parameter $\tilde{B}$. It can be combined with Proposition 3.26 and the results of Section 3.2.1 to provide guarantees on the output of the Lloyd type algorithm [Chazal et al., 2021, Algorithm 1] (or Algorithm 1), combined with a suitable kernel. We choose to give results for the theoretical kernel $\psi_0 : x \mapsto (1 - ((x-1)\vee 0))\vee 0$, and for the kernel used in Royer et al. [2021], $\psi_{AT} = x \mapsto \exp(-x)$.

Corollary 3.30 :

Assume that $\mathbb{E}(X)$ satisfies a margin condition, and that the requirements of Proposition 3.29 are satisfied. For short, denote by v_i the vectorization of X_i based on the output of Algorithm 1. Then, with probability larger than

$$1 - \exp\left[-C\left(\frac{nr^2 p_{min}^2}{p^2 M^2 \Lambda^2 k^2 d \log(k)} - \frac{p_{min}^2 B^2 r_0^2}{M^2 \Lambda^4 k^2 d \log(k)}\right)\right] - n \max_{\ell \leq L} N_\ell^{-\left(\frac{(\kappa \tilde{B})^{d_\ell} f_{min,\ell} N_\ell}{C_\ell d_\ell \log(N_\ell)}\right)},$$

where κ and C are small enough constants, we have

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

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

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

This result can be turned into probability bounds on the exactness of the output of hierarchical clustering schemes applied to the sample points. For instance, on the probability event described by Corollary 3.30, Single Linkage with norm $\|.\|_\infty$ will provide an exact clustering. The probability bound in Corollary 3.30 shed some light on the quality of sampling of each shape that is required to achieve a perfect classification: roughly, for N_ℓ in $\Omega(\log(n))$, the probability of misclassification can be controlled. Provided deviation bounds were also available for Algorithm 2, a similar result might be stated. Note that though the key parameter $\tilde{B}$ is not known, in practice it can be scaled as several times the minimum distance between two points of a diagram.
To determine whether the margin condition holds for $\mathbb{E}(X)$ is satisfied in the mixture of shapes framework is not straightforward. As mentioned below Theorem 3.21, it is likely that milder condition on the regularity of $\mathbb{E}(X)$ could be enough to guarantee the convergence of quantization algorithms. Interestingly, regularity results on $\mathbb{E}(X)$ are derived in Chazal and Divol [2019]. Though purely conjectural at the moment, the intuition that our quantization/vectorization scheme works for this 2-dimensional particular instance of sample measures is supported by the experimental results exposed in the following section.

3.2.3.2 Large-scale graph classification

We intend to apply our quantization/vectorization method to the large-scale graph classification problem. Provided that graphs may be considered as measures, our vectorization method provides an embedding into $\mathbb{R}^k$ that may be combined with standard learning algorithms. In this framework, our vectorization procedure can be thought of as a dimensionality reduction technique for supervised learning, that is performed in an unsupervised way. The notion of shattering codebook introduced in Section 3.2.2 is still relevant in this context, since a (p, r, Δ)-shattering codebook would lead to exact sample classification if combined with a classification tree for instance.
We provide results for the Atol procedure Royer et al. [2021], that is a particular instance of our general scheme with kernel $\psi = \psi_{AT} = x \mapsto e^{-x}$. As exposed above, since Atol is an unsupervised procedure that is not specifically conceived for graphs, it is likely that other dedicated techniques would provide better performances in this graph

method		SF	NetLSD	FGSD	GeoScat	ATOL
problem	size	de Lara and Pineau [2018]	Tsitsulin et al. [2018]	Verma and Zhang [2017]	Gao et al. [2019]	
`reddit threads`	(203K)	81.4±.2	**82.7±.1**	82.5±.2	80.0±.1	80.7±.1
`twitch egos`	(127K)	67.8±.3	63.1±.2	**70.5±.3**	69.7±.1	69.7±.1
`github stargazers`	(12.7K)	55.8±.1	63.2±.1	65.6±.1	54.6±.3	**72.3±.4**
`deezer ego nets`	(9.6K)	50.1±.1	52.2±.1	**52.6±.1**	52.2±.3	51.0±.6

Table 3.3: Large graph binary classification problems. Mean ROC-AUC and standard deviations.

classification problem. However, its simplicity can be a competitive advantage for large-scale applications.

There are multiple ways to interpret graphs as measures. In this section we borrow from Carrière et al. [2019]: for a diffusion time $t > 0$ we compute the Heat Kernel Signatures (HKS_t) for all vertices in a graph, so that each graph $G(V, E)$ is embedded in $\mathbb{R}^{|V|}$ (see, e.g., [Carrière et al., 2019, Section 2.2]). Then four graph descriptors per diffusion time may be computed using the extended persistence framework (see, e.g., [Carrière et al., 2019, Section 2.1]), that is four types of persistence diagrams (PDs). Schematically for a graph $G(V, E)$ the descriptors are derived as:

$$\begin{aligned} G(V,E) &\xrightarrow[\text{signatures}]{\text{heat kernel}} \text{HKS}_t(G) \in \mathbb{R}^{|V|}, \\ G(V,E) &\xrightarrow[\text{persistence}]{\text{extended}} \text{PD}(\text{HKS}_t(G), G) \in (\mathcal{M}(\mathbb{R}^2))^4. \end{aligned} \tag{3.7}$$

In these experiments we will show that the graph embedding strategy of (3.7) paired with ATOL can perform up to the state of the art on large-scale graph classification problems.

Large-scale binary classification from Rozemberczki et al. [2020]

Recently Rozemberczki et al. [2020] introduced large-scale graph datasets of social or web origin. For each dataset the associated task is binary classification. The authors perform a 80% train/test split of the data and report mean area under the curve (AUC) along with standard errors over a hundred experiments for all the following graph embedding methods. SF from de Lara and Pineau [2018] is a simple graph embedding method that extracts the k lowest spectral values from the graph Laplacian, and a standard Random Forest Classifier (RFC) for classification. NetLSD from Tsitsulin et al. [2018] uses a more refined representation of the graph Laplacian, the heat trace signature (a global variant of the HKS) of a graph using 250 diffusion times, and a 1-layer neural network (NN) for classification. FGSD from Verma and Zhang [2017] computes the biharmonic spectral distance of a graph and uses histogram vectorization with small binwidths that results in vectorization of size $[100, 1000000]$, with a Support Vector Machine (SVM) classifier. GeoScattering from Gao et al. [2019] uses graph wavelet transform to produce 125 graph embedding features, also with a SVM classifier.

We add our own results for ATOL paired with (3.7): we use the extended persistence diagrams as input for ATOL computed from the HKS values with diffusion times $t_1 = .1, t_2 = 10$, and vectorize the diagrams with budget $k = 10$ for each diagram type and diffusion time so that the resulting vectorization for graph G is $v_{\text{ATOL}}(G) \in \mathbb{R}^{2\times4\times10}$. We then train a standard RFC (as in de Lara and Pineau [2018], we use the implementation from `sklearn` Pedregosa et al. [2011] with all default parameters) on the resulting vectorized measures.

The results are shown Table 3.3. ATOL is close to or over the state-of-the-art for all four datasets, for a much lighter computational cost. Most of these methods operate directly from graph Laplacians so they are fairly comparable, in essence or as for the dimension of the embedding that is used. The most positive results on `github stargazers` improves on the best method by more than 6 points. Interestingly, there are other ways than (3.7) to represent graphs by measure, for which our methodology readily applies. We present below a short example.

A variant of the large-scale graph classification from Royer et al. [2021]
The graph classification tasks above were binary classifications. Yanardag and Vishwanathan [2015] introduced now popular datasets of large-scale graphs associated with multiple classes. These datasets have been tackled with top performant graph methods including the graph kernel methods RetGK from Zhang et al. [2018], WKPI from Zhao and Wang [2019] and GNTK from Du et al. [2019] (combined with a graph neural network), and the aforementioned graph embedding method FGSD from Verma and Zhang [2017]. PersLay from Carrière et al. [2019] is not designed for graphs but used (3.7) as input for a 1-NN classifier. Lastly, in Royer et al. [2021], we also used (3.7) and the same diagrams as input data for the Atol procedure and obtained competitive results within reasonable computation times (the largest is 110 seconds, for the `COLLAB` dataset, see [Royer et al., 2021, Section 3.1] for all computation times).
Here we propose to bypass the topological feature extraction step in (3.7). Instead of topological descriptors, some simpler graph descriptors are preferred: we compute four HKS descriptors corresponding to diffusion times $t_1 = .1, t_2 = 1, t_3 = 10, t_4 = 100$ for all vertices in a graph, but this time directly interpret the output as a measure embedding in dimension 4. From there we use Atol with $k = 80$ budget. Therefore each graph $G(V, E)$ is embedded in $\mathbb{R}^{4|V|}$ seen as $\mathcal{M}(\mathbb{R}^4)$ and our measure vectorization framework is readily applicable from there. To sum-up we now use the point of view:

$$G(V, E) \xrightarrow[\text{signatures}]{\text{heat kernel}} \mathrm{HKS}_{t_1,t_2,t_3,t_4}(G) \in \mathbb{R}^{4|V|} \approx \mathcal{M}(\mathbb{R}^4). \tag{3.8}$$

It is important to note that the proposed transformation from graphs to measures can be computed using any type of node or edge embedding. Our vectorization method could be applied the same way as for HKS's.

method	RetGK	FGSD	WKPI	GNTK	PersLay	Atol	Atol
problem	Zhang et al. [2018]	Verma and Zhang [2017]	Zhao and Wang [2019]	Du et al. [2019]	Carrière et al. [2019]	with (3.7)	with (3.8)
`REDDIT` (5K, 5 classes)	56.1±.5	47.8	59.5±.6	—	55.6±.3	**67.1±.3**	**66.1±.2**
`REDDIT` (12K, 11 classes)	48.7±.2	—	48.5±.5	—	47.7±.2	**51.4±.2**	**50.7±.3**
`COLLAB` (5K, 3 classes)	81.0±.3	80.0	—	83.6±.1	76.4±.4	**88.3±.2**	**88.5±.1**
`IMDB-B` (1K, 2 classes)	71.9±1.	73.6	75.1±1.1	**76.9±3.6**	71.2±.7	74.8±.3	73.9±.5
`IMDB-M` (1.5K, 3 classes)	47.7±.3	**52.4**	48.4±.5	**52.8±4.6**	48.8±.6	47.8±.7	47.0±.5

Table 3.4: Mean accuracies and standard deviations for the large multi-classes problems of Yanardag and Vishwanathan [2015].

On Table 3.4 we quote results and competitors from Royer et al. [2021] and on the right column we add our own experiment with Atol. The Atol methodology works very efficiently with the direct HKS embedding as well, although the results tend to be slightly inferior. This may hint at the fact that although persistence diagrams are not essential to capturing signal from this dataset, they can be a significant addition for doing so. Overall Atol with (3.8), though much lighter than Atol with (3.7), remains competitive for large-scale graph multiclass classification.
Let us conclude this experimental section by recalling that since our vectorization scheme is fully unsupervised, its advantages mostly rely in its simplicity (hence reduced computational cost) and large scope of applications. Some task-dedicated embedding achieve better overall performances, that is not surprising. This empirical study shows that on the whole our vectorization scheme fairly compares with dedicated methods, for a much lighter computational cost. This computational benefit is particularly prominent in large dimensions, an instance of this point is given for a text classification task in [Chazal et al., 2021, Section 4.3].

3.3 Discussion and directions for future research

Dimensionality reduction for persistence diagram classification

The quantization-based vectorization of persistence diagrams proposed in Section 3.2.3, though unsupervised, allows for efficient classification whenever a code point takes place in a discriminative area. Then, a natural question is to find such a discriminative area, whenever the sample is labeled as $X_1, \ldots, X_{N_1}, Y_1, \ldots, Y_{N_2}$, with $X_i \sim X$ and $Y_i \sim Y$, or slightly generalizing, a function $f : \mathbb{R}^2 \to \mathbb{R}^k$ such that $\|X(du)(f(u)) - Y(du)(f(u))\|$ is as large as possible with high probability. Through this lens, a classification-adapted vectorization of measures is $X(f) = (X(du)(f_1(u)), \ldots, X(du)(f_k(u)))$, for such a discriminative f with coordinates f_j.
Interestingly, several persistence diagram vectorization schemes are indeed of the form $X(f)$, where f has some particular structure. For instance, *persistence silhouettes* (Chazal et al. [2015b]) evaluated at some $t > 0$ is just $X \mapsto X(du)\phi_t(u)$, where $\phi_t(u) = (u^{(2)} - |t - u^{(1)}|)_+$ (in a birth/persistence representation of a diagram). Similarly, *persistence image* Adams et al. [2017] evaluates, at each pixel p of a pixel grid, the functional $X \mapsto \int_p (X \star g_\sigma)(x)dx$ (mass given by a Gaussian convolution to a pixel), that can be written as $X \mapsto X(du)(f_p(u))$ for some continuous and compactly supported f_p, using the Riesz representation Theorem.
Dimensionality reduction in such an integral vectorization framework might be performed in two ways. First, using standard dimensionality reduction techniques for classification based on a prior multidimensional vectorization, such as ℓ_1-penalized logistic regression based on persistence image as in Obayashi et al. [2018]. Second, by constructing discriminative $\hat{f}_1, \ldots, \hat{f}_k$ based on sample, for instance by seeking for a discriminative function between $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ in a suitable RKHS, as suggested by Gretton et al. [2012]. Both of these approaches would lead to interesting developments in measure classification, provided theoretical conditions that can guarantee discrimination are well understood. This is ongoing work with Olympio Hacquard and Gilles Blanchard.

Sufficient conditions for (optimal) convergence of quantization algorithms

The key assumption to prove convergence of quantization algorithms 2 and [Chazal et al., 2021, Algorithm 1] was the margin condition (see Definition 3.19), that is roughly valid whenever the source distribution is concentrated enough around k poles. Though adapted to a clustering framework (see Levrard [2018]), such a condition is not adapted to quantization of uniform distribution on structures for large k's, that are key instances of quantization applications (see Section 2.1). To give an example, easy computation show that such a condition does not hold for the uniform distribution on $[-1, 1]$, for $k = 2$. However, convergence results for Lloyd-type algorithms may be derived under structural assumptions that bypass the margin condition, in dimension $d = 1$: by requiring the source distribution to have a log-concave density (Kieffer [1983]), or to have a continuous and compactly supported density (unpublished result). The basic idea that grounds these two results is that iterations of the Lloyd algorithm can only move codepoints in directions where co-coercitivity occurs (see Lemma 3.20), that is less requiring than co-coercitivity in every direction a margin condition entails. Extension of these results in arbitrary dimension d is ongoing work.
Up to our knowledge, fast rates of convergence (faster than $1/\sqrt{n}$) for ERM or iterative algorithm outputs are available in the k-means case only, that is Bregman divergence with $\phi(u) = \|u\|^2$. Deriving fast rates for general Bregman quantization might be performed by two ways. First by extending the margin conditions to the Bregman case, that is not trivial since Bregman divergences do not satisfy the triangular inequality that is required in the proof of [Levrard, 2015, Lemma 4.2]. The second option would be to adapt the aforementioned principle, by proving that iterations of Lloyd type algorithms move codepoints in positive eigenspaces of the Hessian matrix of the Bregman distortion function at optimal codebooks (when properly defined). Such a result could also pave the way to prove optimality of mini-batches version of Algorithm 1 in the Bregman case.

Bibliography

E. Aamari. *Convergence Rates for Geometric Inference*. Theses, Université Paris-Saclay, Sept. 2017. URL https://hal.inria.fr/tel-01607782.

E. Aamari and C. Levrard. Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction. *Discrete Comput. Geom.*, 59(4):923–971, 2018. ISSN 0179-5376. doi: 10.1007/s00454-017-9962-z. URL https://doi.org/10.1007/s00454-017-9962-z.

E. Aamari and C. Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *Ann. Statist.*, 47(1):177–204, 2019. ISSN 0090-5364. doi: 10.1214/18-AOS1685. URL https://doi.org/10.1214/18-AOS1685.

E. Aamari, J. Kim, F. Chazal, B. Michel, A. Rinaldo, and L. Wasserman. Estimating the reach of a manifold. *Electron. J. Stat.*, 13(1):1359–1399, 2019. ISSN 1935-7524. doi: 10.1214/19-ejs1551. URL https://doi.org/10.1214/19-ejs1551.

E. Aamari, C. Aaron, and C. Levrard. Minimax boundary estimation and estimation with boundary. 2021.

C. Aaron and O. Bodart. Local convex hull support and boundary estimation. *J. Multivariate Anal.*, 147:82–101, 2016. ISSN 0047-259X. URL https://doi.org/10.1016/j.jmva.2016.01.003.

C. Aaron and A. Cholaquidis. On boundary detection. *Ann. Inst. Henri Poincaré Probab. Stat.*, 56(3):2028–2050, 2020. ISSN 0246-0203. doi: 10.1214/19-AIHP1027. URL https://doi.org/10.1214/19-AIHP1027.

C. Aaron, A. Cholaquidis, and R. Fraiman. Surface and length estimation based on Crofton's formula. working paper or preprint, Feb. 2021. URL https://hal.archives-ouvertes.fr/hal-02907297.

H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier. Persistence images: a stable vector representation of persistent homology. *J. Mach. Learn. Res.*, 18:Paper No. 8, 35, 2017. ISSN 1532-4435.

P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. Geometric approximation via coresets. In *Combinatorial and computational geometry*, volume 52 of *Math. Sci. Res. Inst. Publ.*, pages 1–30. Cambridge Univ. Press, Cambridge, 2005. doi: 10.4171/PRIMS/172. URL https://doi.org/10.4171/PRIMS/172.

Y. Aizenbud and B. Sober. Non-parametric estimation of manifolds from noisy data, 2021.

S. B. Alexander and R. L. Bishop. Gauss equation and injectivity radii for subspaces in spaces of curvature bounded above. *Geom. Dedicata*, 117:65–84, 2006. ISSN

0046-5755. doi: 10.1007/s10711-005-9011-6. URL http://dx.doi.org/10.1007/s10711-005-9011-6.

A. Antos. Improved minimax bounds on the test and training distortion of empirically designed vector quantizers. *IEEE Trans. Inform. Theory*, 51(11):4022–4032, 2005. ISSN 0018-9448. doi: 10.1109/TIT.2005.856980. URL https://doi.org/10.1109/TIT.2005.856980.

A. Antos, L. Györfi, and A. György. Individual convergence rates in empirical vector quantizer design. *IEEE Trans. Inform. Theory*, 51(11):4013–4022, 2005. ISSN 0018-9448. doi: 10.1109/TIT.2005.856976. URL https://doi.org/10.1109/TIT.2005.856976.

D. Arthur and S. Vassilvitskii. `k-means++`: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035. ACM, New York, 2007.

D. Attali, A. Lieutier, and D. Salinas. Vietoris–rips complexes also provide topologically correct reconstructions of sampled shapes. *Computational Geometry*, 46(4):448–465, 2013. ISSN 0925-7721. doi: https://doi.org/10.1016/j.comgeo.2012.02.009. URL https://www.sciencedirect.com/science/article/pii/S0925772112001423. 27th Annual Symposium on Computational Geometry (SoCG 2011).

O. Bachem, M. Lucic, S. H. Hassani, and A. Krause. Uniform deviation bounds for k-means clustering. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 283–291. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/bachem17a.html.

A. Baddeley. Integrals on a moving manifold and geometrical probability. *Advances in Appl. Probability*, 9(3):588–603, 1977. ISSN 0001-8678. doi: 10.2307/1426116. URL https://doi.org/10.2307/1426116.

A. Banerjee and R. N. Davé. Robust clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2, 2012.

A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Trans. Inform. Theory*, 51(7):2664–2669, 2005a. ISSN 0018-9448. doi: 10.1109/TIT.2005.850145. URL https://doi.org/10.1109/TIT.2005.850145.

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *J. Mach. Learn. Res.*, 6:1705–1749, 2005b. ISSN 1532-4435.

I. Bárány. Intrinsic volumes and f-vectors of random polytopes. *Math. Ann.*, 285(4): 671–699, 1989. ISSN 0025-5831. doi: 10.1007/BF01452053. URL https://doi.org/10.1007/BF01452053.

P. L. Bartlett, T. Linder, and G. Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Trans. Inform. Theory*, 44(5):1802–1813, 1998. ISSN 0018-9448. doi: 10.1109/18.705560. URL https://doi.org/10.1109/18.705560.

A.-H. Bateni and A. S. Dalalyan. Confidence regions and minimax rates in outlier-robust estimation on the probability simplex. *Electron. J. Stat.*, 14(2):2653–2677, 2020. doi: 10.1214/20-EJS1731. URL https://doi.org/10.1214/20-EJS1731.

C. Berenfeld, J. Harvey, M. Hoffmann, and K. Shankar. Estimating the reach of a manifold via its convexity defect function. *arXiv e-prints*, art. arXiv:2001.08006, Jan. 2020.

G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Trans. Inform. Theory*, 54(2):781–790, 2008. ISSN 0018-9448. doi: 10.1109/TIT.2007.913516. URL https://doi.org/10.1109/TIT.2007.913516.

J.-D. Boissonnat and A. Ghosh. Manifold reconstruction using tangential Delaunay complexes. *Discrete Comput. Geom.*, 51(1):221–267, 2014. ISSN 0179-5376. doi: 10.1007/s00454-013-9557-2. URL https://doi.org/10.1007/s00454-013-9557-2.

J.-D. Boissonnat, L. J. Guibas, and S. Y. Oudot. Manifold reconstruction in arbitrary dimensions using witness complexes. *Discrete Comput. Geom.*, 42(1):37–70, 2009. ISSN 0179-5376. doi: 10.1007/s00454-009-9175-1. URL http://dx.doi.org/10.1007/s00454-009-9175-1.

J.-D. Boissonnat, F. Chazal, and M. Yvinec. *Geometric and topological inference*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2018. ISBN 978-1-108-41089-2; 978-1-108-41939-0. doi: 10.1017/9781108297806. URL https://doi.org/10.1017/9781108297806.

L. Bottou and Y. Bengio. Convergence properties of the k-means algorithms. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1995. URL https://proceedings.neurips.cc/paper/1994/file/a1140a3d0df1c81e24ae954d935e8926-Paper.pdf.

C. Brécheteau and C. Levrard. A k-points-based distance for robust geometric inference. *Bernoulli*, 26(4):3017–3050, 2020. ISSN 1350-7265. doi: 10.3150/20-BEJ1214. URL https://doi.org/10.3150/20-BEJ1214.

C. Brécheteau, A. Fischer, and C. Levrard. Robust Bregman clustering. *Ann. Statist.*, 49 (3):1679–1701, 2021. ISSN 0090-5364. doi: 10.1214/20-aos2018. URL https://doi.org/10.1214/20-aos2018.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984. ISBN 0-534-98053-8; 0-534-98054-6.

C. Brownlees, E. Joly, and G. Lugosi. Empirical risk minimization for heavy-tailed losses. *Ann. Statist.*, 43(6):2507–2536, 2015. ISSN 0090-5364. doi: 10.1214/15-AOS1350. URL https://doi.org/10.1214/15-AOS1350.

C. Brunet-Saumard, E. Genetay, and A. Saumard. K-bmom: a robust lloyd-type clustering algorithm based on bootstrap median-of-means, 2020.

C. Brunet-Saumard, E. Genetay, and A. Saumard. K-bMOM: A robust Lloyd-type clustering algorithm based on bootstrap median-of-means. *Comput. Statist. Data Anal.*, 167:Paper No. 107370, 2022. ISSN 0167-9473. doi: 10.1016/j.csda.2021.107370. URL https://doi.org/10.1016/j.csda.2021.107370.

P. Bubenik. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16(1):77–102, jan 2015. ISSN 1532-4435.

M. Buchet, F. Chazal, S. Y. Oudot, and D. R. Sheehy. Efficient and robust persistent homology for measures. *Comput. Geom.*, 58:70–96, 2016. ISSN 0925-7721. doi: 10.1016/j.comgeo.2016.07.001. URL https://doi.org/10.1016/j.comgeo.2016.07.001.

G. Canas, T. Poggio, and L. Rosasco. Learning manifolds with k-means and k-flats. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/b20bb95ab626d93fd976af958fbc61ba-Paper.pdf.

H. Cardot, P. Cénac, and J.-M. Monnez. A fast and recursive algorithm for clustering large datasets with k-medians. *Comput. Statist. Data Anal.*, 56(6):1434–1449, 2012. ISSN 0167-9473. doi: 10.1016/j.csda.2011.11.019. URL https://doi.org/10.1016/j.csda.2011.11.019.

H. Cardot, P. Cenac, and P.-A. Zitt. Efficient and fast estimation of the geometric median in hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19:18–43, 2013.

M. Carrière, M. Cuturi, and S. Oudot. Sliced Wasserstein kernel for persistence diagrams. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 664–673. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/carriere17a.html.

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

O. Catoni and I. Giulini. Dimension-free PAC-Bayesian bounds for the estimation of the mean of a random vector. *ArXiv e-prints*, Feb. 2018.

F. Cazals and M. Pouget. Estimating differential quantities using polynomial fitting of osculating jets. *Comput. Aided Geom. Design*, 22(2):121–146, 2005. ISSN 0167-8396. doi: 10.1016/j.cagd.2004.09.004. URL https://doi.org/10.1016/j.cagd.2004.09.004.

G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Comput. Statist. Data Anal.*, 14(3):315–332, 1992. ISSN 0167-9473. doi: 10.1016/0167-9473(92)90042-E. URL https://doi.org/10.1016/0167-9473(92)90042-E.

F. Chazal and V. Divol. The density of expected persistence diagrams and its kernel based estimation. *J. Comput. Geom.*, 10(2):127–153, 2019.

F. Chazal and S. Y. Oudot. Towards persistence-based reconstruction in Euclidean spaces. In *Computational geometry (SCG'08)*, pages 232–241. ACM, New York, 2008. doi: 10.1145/1377676.1377719. URL https://doi.org/10.1145/1377676.1377719.

F. Chazal, D. Cohen-Steiner, L. J. Guibas, F. Mémoli, and S. Y. Oudot. Gromov-hausdorff stable signatures for shapes using persistence. *Computer Graphics Forum*, 28(5):1393–1403, 2009. doi: 10.1111/j.1467-8659.2009.01516.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2009.01516.x.

F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for probability measures. *Found. Comput. Math.*, 11(6):733–751, 2011. ISSN 1615-3375. doi: 10.1007/s10208-011-9098-0. URL https://doi.org/10.1007/s10208-011-9098-0.

F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba. Persistence-based clustering in Riemannian manifolds. *J. ACM*, 60(6):Art. 41, 38, 2013. ISSN 0004-5411. doi: 10.1145/2535927. URL https://doi.org/10.1145/2535927.

F. Chazal, B. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman. Subsampling methods for persistent homology. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2143–2151, Lille, France, 07–09 Jul 2015a. PMLR. URL http://proceedings.mlr.press/v37/chazal15.html.

F. Chazal, B. T. Fasy, F. Lecci, A. Rinaldo, and L. Wasserman. Stochastic convergence of persistence landscapes and silhouettes. *J. Comput. Geom.*, 6(2):140–161, 2015b.

F. Chazal, M. Glisse, C. Labruère, and B. Michel. Convergence rates for persistence diagram estimation in topological data analysis. *J. Mach. Learn. Res.*, 16:3603–3635, 2015c. ISSN 1532-4435.

F. Chazal, M. Glisse, C. Labruère, and B. Michel. Convergence rates for persistence diagram estimation in topological data analysis. *J. Mach. Learn. Res.*, 16:3603–3635, 2015d. ISSN 1532-4435.

F. Chazal, P. Massart, and B. Michel. Rates of convergence for robust geometric inference. *Electron. J. Stat.*, 10(2):2243–2286, 2016. doi: 10.1214/16-EJS1161. URL https://doi.org/10.1214/16-EJS1161.

F. Chazal, C. Levrard, and M. Royer. Clustering of measures via mean measure quantization. *Electron. J. Stat.*, 15(1):2060–2104, 2021. doi: 10.1214/21-ejs1834. URL https://doi.org/10.1214/21-ejs1834.

Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8):790–799, Aug. 1995. ISSN 0162-8828. doi: 10.1109/34.400568. URL https://doi.org/10.1109/34.400568.

K. L. Clarkson. Building triangulations using ϵ-nets (extended abstract). In *STOC'06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 326–335. ACM, New York, 2006. doi: 10.1145/1132516.1132564. URL https://doi.org/10.1145/1132516.1132564.

D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete Comput. Geom.*, 37(1):103–120, 2007. ISSN 0179-5376. doi: 10.1007/s00454-006-1276-5. URL https://doi.org/10.1007/s00454-006-1276-5.

J. A. Cuesta-Albertos, A. Gordaliza, and C. Matrán. Trimmed k-means: an attempt to robustify quantizers. *Ann. Statist.*, 25(2):553–576, 04 1997. doi: 10.1214/aos/1031833664. URL http://dx.doi.org/10.1214/aos/1031833664.

A. Cuevas. Set estimation: another bridge between statistics and geometry. *Bol. Estad. Investig. Oper.*, 25(2):71–85, 2009. ISSN 1889-3805.

A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *Ann. Statist.*, 25 (6):2300–2312, 1997. ISSN 0090-5364. doi: 10.1214/aos/1030741073. URL https://doi.org/10.1214/aos/1030741073.

A. Cuevas and A. Rodríguez-Casal. On boundary estimation. *Adv. in Appl. Probab.*, 36 (2):340–354, 2004. ISSN 0001-8678. doi: 10.1239/aap/1086957575. URL http://dx.doi.org/10.1239/aap/1086957575.

A. Cuevas, M. Febrero, and R. Fraiman. Cluster analysis: a further approach based on density estimation. *Comput. Statist. Data Anal.*, 36(4):441–459, 2001. ISSN 0167-9473. doi: 10.1016/S0167-9473(00)00052-9. URL https://doi.org/10.1016/S0167-9473(00)00052-9.

A. Cuevas, R. Fraiman, and B. Pateiro-López. On statistical properties of sets fulfilling rolling-type conditions. *Adv. in Appl. Probab.*, 44(2):311–329, 2012. ISSN 0001-8678. doi: 10.1239/aap/1339878713. URL https://doi.org/10.1239/aap/1339878713.

M. Cuturi and A. Doucet. Fast Computation of Wasserstein Barycenters. *arXiv e-prints*, art. arXiv:1310.4375, Oct 2013.

N. de Lara and E. Pineau. A simple baseline algorithm for graph classification, 2018.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977. ISSN 0035-9246. URL http://links.jstor.org/sici?sici=0035-9246(1977)39:1<1:MLFIDV>2.0.CO;2-Z&origin=MSN. With discussion.

L. Devroye and G. L. Wise. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J. Appl. Math.*, 38(3):480–488, 1980. ISSN 0036-1399. doi: 10.1137/0138038. URL https://doi.org/10.1137/0138038.

T. K. Dey. *Curve and surface reconstruction: algorithms with mathematical analysis*, volume 23 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2007. ISBN 978-0-521-86370-4; 0-521-86370-8.

T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71, 1997. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(96)00034-3. URL https://www.sciencedirect.com/science/article/pii/S0004370296000343.

P. J. Diggle. A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 153(3):349–362, 1990. doi: 10.2307/2982977. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2982977.

V. Divol. Minimax adaptive estimation in manifold inference, 2020.

V. Divol. Minimax adaptive estimation in manifold inference. working paper or preprint, Oct. 2021. URL https://hal.inria.fr/hal-02440881.

M. P. a. do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser Boston, Inc., Boston, MA, 1992. ISBN 0-8176-3490-8. doi: 10.1007/978-1-4757-2201-7. URL https://doi.org/10.1007/978-1-4757-2201-7. Translated from the second Portuguese edition by Francis Flaherty.

D. Donoho and P. J. Huber. The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pages 157–184. Wadsworth, Belmont, CA, 1983.

S. S. Du, K. Hou, R. R. Salakhutdinov, B. Poczos, R. Wang, and K. Xu. Graph neural tangent kernel: Fusing graph neural networks with graph kernels. In *Advances in Neural Information Processing Systems 32*, pages 5724–5734. Curran Associates, Inc., 2019.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. Wiley-Interscience, New York, second edition, 2001. ISBN 0-471-05669-3.

L. Dümbgen and G. Walther. Rates of convergence for random approximations of convex sets. *Adv. in Appl. Probab.*, 28(2):384–393, 1996. ISSN 0001-8678. doi: 10.2307/1428063. URL http://dx.doi.org/10.2307/1428063.

H. Edelsbrunner. Weighted alpha shapes. Technical report, Champaign, IL, USA, 1992.

H. Edelsbrunner and J. L. Harer. *Computational topology*. American Mathematical Society, Providence, RI, 2010. ISBN 978-0-8218-4925-5. An introduction.

H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. volume 28, pages 511–533. 2002. doi: 10.1007/s00454-002-2885-2. URL https://doi.org/10.1007/s00454-002-2885-2. Discrete and computational geometry and graph drawing (Columbia, SC, 2001).

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

S. Evert. A simple lnre model for random character sequences. In *In Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles (Louvain-la-Neuve*, pages 411–422, 2004.

J. Fan and Y. K. Truong. Nonparametric regression with errors in variables. *Ann. Statist.*, 21(4):1900–1925, 1993. ISSN 0090-5364. doi: 10.1214/aos/1176349402. URL https://doi.org/10.1214/aos/1176349402.

B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh. Confidence sets for persistence diagrams. *Ann. Statist.*, 42(6):2301–2339, 2014. ISSN 0090-5364. doi: 10.1214/14-AOS1252. URL https://doi.org/10.1214/14-AOS1252.

H. Federer. Curvature measures. *Trans. Amer. Math. Soc.*, 93:418–491, 1959. ISSN 0002-9947. doi: 10.2307/1993504. URL https://doi.org/10.2307/1993504.

H. Federer. *Geometric measure theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York, 1969.

C. Fefferman, S. Ivanov, Y. Kurylev, M. Lassas, and H. Narayanan. Fitting a putative manifold to noisy data. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 688–720. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/fefferman18a.html.

A. Fischer. Quantization and clustering with Bregman divergences. *J. Multivariate Anal.*, 101(9):2207–2221, 2010. ISSN 0047-259X. doi: 10.1016/j.jmva.2010.05.008. URL https://doi.org/10.1016/j.jmva.2010.05.008.

H. Fritz, L. A. Garcia-Escudero, and A. Mayo-Iscar. tclust: An R package for a trimming approach to cluster analysis. *Journal of Statistical Software*, 47(12):1–26, 2012. URL http://www.jstatsoft.org/v47/i12/.

S. Gallot, D. Hulin, and J. Lafontaine. *Riemannian geometry*. Universitext. Springer-Verlag, Berlin, third edition, 2004. ISBN 3-540-20493-8. doi: 10.1007/978-3-642-18855-8. URL https://doi.org/10.1007/978-3-642-18855-8.

F. Gao, G. Wolf, and M. Hirn. Geometric scattering for graph data analysis. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2122–2131, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/gao19e.html.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. *Ann. Statist.*, 40(2): 941–963, 2012a. ISSN 0090-5364. doi: 10.1214/12-AOS994. URL https://doi.org/10.1214/12-AOS994.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Minimax manifold estimation. *J. Mach. Learn. Res.*, 13:1263–1291, 2012b. ISSN 1532-4435.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Nonparametric ridge estimation. *Ann. Statist.*, 42(4):1511–1545, 2014. ISSN 0090-5364. doi: 10.1214/14-AOS1218. URL https://doi.org/10.1214/14-AOS1218.

A. Gersho and R. M. Gray. *Vector quantization and signal compression*. Kluwer Academic Publishers, Norwell, MA, USA, 1991. ISBN 0-7923-9181-0.

T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoret. Comput. Sci.*, 38(2-3):293–306, 1985. ISSN 0304-3975. doi: 10.1016/0304-3975(85)90224-5. URL https://doi.org/10.1016/0304-3975(85)90224-5.

A. Gordaliza. Best approximations to random variables based on trimming procedures. *J. Approx. Theory*, 64(2):162–180, 1991. ISSN 0021-9045. doi: 10.1016/0021-9045(91)90072-I. URL https://doi.org/10.1016/0021-9045(91)90072-I.

J. C. Gower and G. J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Appl. Statist.*, 18:54–61, 1969. ISSN 0035-9254. doi: 10.2307/2346439. URL https://doi.org/10.2307/2346439.

S. Graf and H. Luschgy. *Foundations of quantization for probability distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000. ISBN 3-540-67394-6. doi: 10.1007/BFb0103945. URL https://doi.org/10.1007/BFb0103945.

S. Graf, H. Luschgy, and G. Pagès. Optimal quantizers for Radon random vectors in a Banach space. *J. Approx. Theory*, 144(1):27–53, 2007. ISSN 0021-9045. doi: 10.1016/j.jat.2006.04.006. URL https://doi.org/10.1016/j.jat.2006.04.006.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012. ISSN 1532-4435.

P. M. Gruber. Optimum quantization and its applications. *Adv. Math.*, 186(2):456–497, 2004. ISSN 0001-8708. doi: 10.1016/j.aim.2003.07.017. URL https://doi.org/10.1016/j.aim.2003.07.017.

L. Guibas, D. Morozov, and Q. Mérigot. Witnessed k-distance. *Discrete Comput. Geom.*, 49(1):22–45, 2013. ISSN 0179-5376. doi: 10.1007/s00454-012-9465-x. URL https://doi.org/10.1007/s00454-012-9465-x.

P. Hartman. On geodesic coordinates. *Amer. J. Math.*, 73:949–954, 1951. ISSN 0002-9327. doi: 10.2307/2372125. URL https://doi.org/10.2307/2372125.

A. Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge, 2002. ISBN 0-521-79160-X; 0-521-79540-0.

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

P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101, 1964. ISSN 0003-4851. doi: 10.1214/aoms/1177703732. URL https://doi.org/10.1214/aoms/1177703732.

P. J. Huber. A robust version of the probability ratio test. *Ann. Math. Statist.*, 36: 1753–1758, 1965. ISSN 0003-4851. doi: 10.1214/aoms/1177699803. URL https://doi.org/10.1214/aoms/1177699803.

L. Kaufman and P. J. Rousseeuw. *Finding groups in data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1990. ISBN 0-471-87876-6. doi: 10.1002/9780470316801. URL https://doi.org/10.1002/9780470316801. An introduction to cluster analysis, A Wiley-Interscience Publication.

J. C. Kieffer. Uniqueness of locally optimal quantizer for log-concave density and convex error weighting function. *IEEE Trans. Inform. Theory*, 29(1):42–47, 1983. ISSN 0018-9448. doi: 10.1109/TIT.1983.1056622. URL https://doi.org/10.1109/TIT.1983.1056622.

A. K. H. Kim and H. H. Zhou. Tight minimax rates for manifold estimation under Hausdorff loss. *Electron. J. Stat.*, 9(1):1562–1582, 2015. ISSN 1935-7524. doi: 10.1214/15-EJS1039. URL http://dx.doi.org/10.1214/15-EJS1039.

Y. Klochkov, A. Kroshnin, and N. Zhivotovskiy. Robust k-means clustering for distributions with two moments, 2020.

A. P. Korostelëv and A. B. Tsybakov. Asymptotic efficiency of the estimation of a convex set. *Problemy Peredachi Informatsii*, 30(4):33–44, 1994. ISSN 0555-2923.

A. Kumar and R. Kannan. Clustering with spectral norm and the k-means algorithm. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science—FOCS 2010*, pages 299–308. IEEE Computer Soc., Los Alamitos, CA, 2010.

G. Lecué and M. Lerasle. Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 48(2):906 – 931, 2020. doi: 10.1214/19-AOS1828. URL https://doi.org/10.1214/19-AOS1828.

J. M. Lee. *Introduction to topological manifolds*, volume 202 of *Graduate Texts in Mathematics*. Springer, New York, second edition, 2011. ISBN 978-1-4419-7939-1. doi: 10.1007/978-1-4419-7940-7. URL https://doi.org/10.1007/978-1-4419-7940-7.

J. W. Leis. *Quantization and Coding*, pages 363–451. 2019. doi: 10.1002/9781119470663.ch5.

C. Levrard. Fast rates for empirical vector quantization. *Electron. J. Stat.*, 7:1716–1746, 2013. doi: 10.1214/13-EJS822. URL https://doi.org/10.1214/13-EJS822.

C. Levrard. *Quantification vectorielle en grande dimension : vitesses de convergence et sélection de variables*. PhD thesis, 2014. URL http://www.theses.fr/2014PA112214. Thèse de doctorat dirigée par Massart, Pascal et Biau, Gérard Mathématiques Paris 11 2014.

C. Levrard. Nonasymptotic bounds for vector quantization in Hilbert spaces. *Ann. Statist.*, 43(2):592–619, 2015. ISSN 0090-5364. doi: 10.1214/14-AOS1293. URL https://doi.org/10.1214/14-AOS1293.

C. Levrard. Quantization/clustering: when and why does k-means work? *J. SFdS*, 159 (1):1–26, 2018.

T. Linder. Learning-theoretic methods in vector quantization. In *Principles of nonparametric learning (Udine, 2001)*, volume 434 of *CISM Courses and Lect.*, pages 163–210. Springer, Vienna, 2002.

C. Liu and M. Belkin. Clustering with bregman divergences: An asymptotic analysis. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2351–2359, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

S. P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inform. Theory*, 28(2): 129–137, 1982. ISSN 0018-9448. doi: 10.1109/TIT.1982.1056489. URL https://doi.org/10.1109/TIT.1982.1056489.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)*, pages Vol. I: Statistics, pp. 281–297. Univ. California Press, Berkeley, Calif., 1967.

M. Maggioni, S. Minsker, and N. Strawn. Multiscale dictionary learning: non-asymptotic bounds and robustness. *J. Mach. Learn. Res.*, 17:Paper No. 2, 51, 2016. ISSN 1532-4435.

M. Mahajan, P. Nimbhorkar, and K. Varadarajan. The planar k-means problem is NP-hard. *Theoret. Comput. Sci.*, 442:13–21, 2012. ISSN 0304-3975. doi: 10.1016/j.tcs.2010.05.034. URL https://doi.org/10.1016/j.tcs.2010.05.034.

R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 2019. ISBN 978-1-119-21468-7. Theory and methods (with R), Second edition of [MR2238141].

G. McLachlan and D. Peel. *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York, 2000. ISBN 0-471-00626-2. doi: 10.1002/0471721182. URL https://doi.org/10.1002/0471721182.

J. R. Munkres. *Elements of algebraic topology*. Addison-Wesley Publishing Company, Menlo Park, CA, 1984. ISBN 0-201-04586-9.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 849–856, Cambridge, MA, USA, 2001. MIT Press.

P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, 39(1-3):419–441, 2008. ISSN 0179-5376. doi: 10.1007/s00454-008-9053-2. URL https://doi.org/10.1007/s00454-008-9053-2.

I. Obayashi, Y. Hiraoka, and M. Kimura. Persistence diagrams with linear machine learning models. *J. Appl. Comput. Topol.*, 1(3-4):421–449, 2018. ISSN 2367-1726. doi: 10.1007/s41468-018-0013-5. URL https://doi.org/10.1007/s41468-018-0013-5.

G. Pagès. A space quantization method for numerical integration. *J. Comput. Appl. Math.*, 89(1):1–38, 1998. ISSN 0377-0427. doi: 10.1016/S0377-0427(97)00190-8. URL https://doi.org/10.1016/S0377-0427(97)00190-8.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

D. Pollard. Strong consistency of k-means clustering. *Ann. Statist.*, 9(1):135–140, 1981. ISSN 0090-5364. URL http://links.jstor.org/sici?sici=0090-5364(198101)9:1<135:SCOC>2.0.CO;2-7&origin=MSN.

D. Pollard. A central limit theorem for k-means clustering. *Ann. Probab.*, 10(4):919–926, 1982. ISSN 0091-1798. URL http://links.jstor.org/sici?sici=0091-1798(198211)10:4<919:ACLTFC>2.0.CO;2-Q&origin=MSN.

N. Puchkin and V. Spokoiny. Structure-adaptive manifold estimation. *arXiv e-prints*, art. arXiv:1906.05014, June 2019.

J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision*, pages 435–446, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-24785-9.

I. W. Renner, J. Elith, A. Baddeley, W. Fithian, T. Hastie, S. J. Phillips, G. Popovic, and D. I. Warton. Point process models for presence-only analysis. *Methods in Ecology and Evolution*, 6(4):366–379, 2015. doi: 10.1111/2041-210X.12352. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12352.

A. Rodríguez Casal. Set estimation under convexity type assumptions. *Ann. Inst. H. Poincaré Probab. Statist.*, 43(6):763–774, 2007a. ISSN 0246-0203. doi: 10.1016/j.anihpb.2006.11.001. URL https://doi.org/10.1016/j.anihpb.2006.11.001.

A. Rodríguez Casal. Set estimation under convexity type assumptions. *Ann. Inst. H. Poincaré Probab. Statist.*, 43(6):763–774, 2007b. ISSN 0246-0203. doi: 10.1016/j.anihpb.2006.11.001. URL https://doi.org/10.1016/j.anihpb.2006.11.001.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.

M. Royer, F. Chazal, C. Levrard, Y. Umeda, and Y. Ike. Atol: Measure vectorization for automatic topologically-oriented learning. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1000–1008. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/royer21a.html.

B. Rozemberczki, O. Kiss, and R. Sarkar. An api oriented open-source python framework for unsupervised learning on graphs, 2020.

S. Shirota, A. E. Gelfand, and J. Mateu. Analyzing Car Thefts and Recoveries with Connections to Modeling Origin-Destination Point Patterns. *arXiv e-prints*, art. arXiv:1701.05863, Jan 2017.

A. Singer and H.-T. Wu. Vector diffusion maps and the connection Laplacian. *Comm. Pure Appl. Math.*, 65(8):1067–1144, 2012. ISSN 0010-3640. doi: 10.1002/cpa.21395. URL https://doi.org/10.1002/cpa.21395.

G. Singh, F. Memoli, T. Ishkhanov, G. Sapiro, G. Carlsson, and D. L. Ringach. Topological analysis of population activity in visual cortex. *Journal of Vision*, 8(8): 11–11, 06 2008. ISSN 1534-7362. doi: 10.1167/8.8.11. URL https://doi.org/10.1167/8.8.11.

B. Sober, Y. Aizenbud, and D. Levin. Approximation of functions over manifolds: A moving least-squares approach. *Journal of Computational and Applied Mathematics*, 383:113140, Feb 2021. ISSN 0377-0427. doi: 10.1016/j.cam.2020.113140. URL http://dx.doi.org/10.1016/j.cam.2020.113140.

A. Solé, V. Caselles, G. Sapiro, and F. Arándiga. Morse description and geometric encoding of digital elevation maps. *IEEE Trans. Image Process.*, 13(9):1245–1262, 2004. ISSN 1057-7149. doi: 10.1109/TIP.2004.832864. URL https://doi.org/10.1109/TIP.2004.832864.

C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 1982. ISSN 0090-5364. URL http://links.jstor.org/sici?sici=0090-5364(198212)10:4<1040:OGROCF>2.0.CO;2-2&origin=MSN.

A. Strehl and J. Ghosh. Cluster ensembles - A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

C. Tang and C. Monteleoni. On lloyd's algorithm: New theoretical insights for clustering in practice. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 1280–1289. PMLR, 2016.

L. T. Taylor Arnold. *Humanities Data in R: Exploring Networks, Geospatial Data, Images, and Text*. Quantitative Methods in the Humanities and Social Sciences. Springer International Publishing, 1st ed. 2015 edition, 2015. ISBN 3319207016,978-3-319-20701-8,978-3-319-20702-5.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319, 2000.

A. Tsitsulin, D. Mottin, P. Karras, A. Bronstein, and E. Müller. Netlsd. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Jul 2018. doi: 10.1145/3219819.3219991. URL http://dx.doi.org/10.1145/3219819.3219991.

K. Usevich and I. Markovsky. Optimization on a grassmann manifold with application to system identification. *Automatica*, 50(6):1656 – 1662, 2014. ISSN 0005-1098. doi: http://dx.doi.org/10.1016/j.automatica.2014.04.010. URL http://www.sciencedirect.com/science/article/pii/S000510981400140X.

D. Vandev. A note on the breakdown point of the least median of squares and least trimmed squares estimators. *Statist. Probab. Lett.*, 16(2):117–119, 1993. ISSN 0167-7152. doi: 10.1016/0167-7152(93)90155-C. URL https://doi.org/10.1016/0167-7152(93)90155-C.

S. Verma and Z.-L. Zhang. Hunt for the unique, stable, sparse and fast feature learning on graphs. In *Advances in Neural Information Processing Systems*, pages 88–98, 2017.

P. Yanardag and S. Vishwanathan. Deep graph kernels. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 1365–1374, New York, NY, USA, 2015. Association for Computing

Machinery. ISBN 9781450336642. doi: 10.1145/2783258.2783417. URL https://doi.org/10.1145/2783258.2783417.

B. Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

Z. Zhang, M. Wang, Y. Xiang, Y. Huang, and A. Nehorai. RetGK: Graph Kernels based on Return Probabilities of Random Walks. In *Advances in Neural Information Processing Systems*, pages 3968–3978, 2018.

Q. Zhao and Y. Wang. Learning metrics for persistence-based summaries and applications for graph classification. In *Advances in Neural Information Processing Systems 32*, pages 9855–9866. Curran Associates, Inc., 2019.

A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, 33(2):249–274, 2005. ISSN 0179-5376. doi: 10.1007/s00454-004-1146-y. URL https://doi.org/10.1007/s00454-004-1146-y.