Minimax adaptive estimation in manifold inference

Vincent Divol

Center for Data Science, New York University
Courant Institute of Mathematical Science, New York University
e-mail: vincent.divol@nyu.edu*; url:* vincentdivol.github.io/

Abstract: We focus on the problem of manifold estimation: given a set of observations sampled close to some unknown submanifold M, one wants to recover information about the geometry of M. Minimax estimators which have been proposed so far all depend crucially on the a priori knowledge of parameters quantifying the underlying distribution generating the sample (such as bounds on its density), whereas those quantities will be unknown in practice. Our contribution to the matter is twofold. First, we introduce a one-parameter family of manifold estimators $(\hat{M}_t)_{t \geq 0}$ based on a localized version of convex hulls, and show that for some choice of t, the corresponding estimator is minimax on the class of models of $\mathcal{C}^2$ manifolds introduced in [21]. Second, we propose a completely data-driven selection procedure for the parameter t, leading to a minimax adaptive manifold estimator on this class of models. This selection procedure actually allows us to recover the Hausdorff distance between the set of observations and M, and can therefore be used as a scale parameter in other settings, such as tangent space estimation.

MSC2020 subject classifications: Primary 62G05; secondary 62C20, 68U05.
Keywords and phrases: Geometric inference, manifold estimation, adaptive estimation, Čech complex.

Received June 2020.

1. Introduction

Manifold inference deals with the estimation of geometric quantities in a random setting. Given $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ a set of i.i.d. observations from some law μ on $\mathbb{R}^D$ supported on (or concentrated around) a d-dimensional manifold M, one wants to produce an estimator $\hat{\theta}$ that estimates accurately some quantity $\theta(M)$ related to the geometry of M such as its dimension d [23, 29, 24], its homology groups [31, 11], its tangent spaces [3, 16], or M itself [21, 22, 30, 2, 3, 32]. Consider for instance the problem of estimating the manifold M with respect to the Hausdorff distance d_H. The quality of an estimator $\hat{M}$ with respect to some law μ, called its μ-risk, is given by the average Hausdorff distance d_H between the estimator and M:

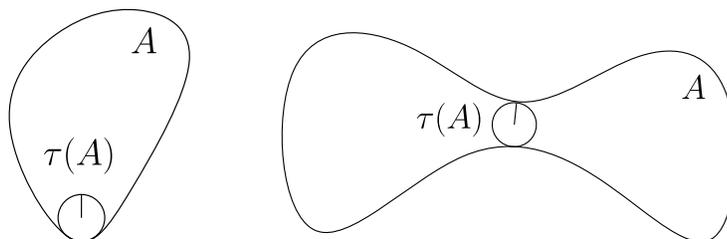$$R_n(\hat{M}, \mu) := \mathbb{E}[d_H(\hat{M}, M)], \tag{1}$$

FIG 1. *If the reach of the curve M is large, then the curve cannot be too pinched (left) and cannot present a tight bottleneck structure (right).*

where $\hat{M} = \hat{M}(\mathcal{X}_n)$ and $\mathcal{X}_n$ is a n-sample of law μ. In reality, the law μ generating the dataset is unknown, and it is more interesting to control the μ-risk uniformly over a set $\mathcal{Q}$ of laws μ, that we call a statistical model. The uniform risk of the estimator $\hat{M}$ on the class $\mathcal{Q}$ is given by,

$$R_n(\hat{M}, \mathcal{Q}) := \sup\{R_n(\hat{M}, \mu) : \ \mu \in \mathcal{Q}\}. \tag{2}$$

while we say that an estimator is *minimax* if it attains (up to a multiplicative constant as n goes to ∞) the *minimax risk*

$$\mathcal{R}_n(\mathcal{Q}) := \inf\{R_n(\hat{M}, \mathcal{Q}) : \ \hat{M} \text{ is an estimator}\}. \tag{3}$$

In geometric inference, several statistical models were introduced, which take into account different noise models and regularities of the manifold M. Let us mention the family of models $\mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}}$ introduced by Genovese et al. in [21], consisting of the laws μ supported on a d-dimensional manifold M satisfying some additional properties. First, we assume that μ has a density f on M, lower bounded by some constant $f_{\min} > 0$ and upper bounded by another constant $f_{\max}$. This ensures that all the parts of the manifold M are approximately evenly sampled: we then say that the law is "almost-uniform" on M. The parameter $\tau_{\min}$ gives a lower bound on the reach $\tau(M)$ of the manifold. The reach is a central notion in geometric inference, defined as the largest radius r such that, if some point x is at distance less than r to M, then there exists a unique projection $\pi_M(x)$ of x on M. As such, it controls both a local regularity of M (a bound on its curvature radius) and a global regularity (namely the presence of a "bottleneck structure"), see also Figure 1.

On the statistical model $\mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}}$, the minimax rate of convergence satisfies

$$c_0 \left(\frac{\ln n}{n}\right)^{2/d} \leq \mathcal{R}_n(\mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}}) \leq c_1 \left(\frac{\ln n}{n}\right)^{2/d}, \tag{4}$$

for two positive constants c_0, c_1 depending on $\tau_{\min}$, $f_{\min}$, $f_{\max}$ and d. The lower bound in this inequality was shown by Kim and Zhou [25], while the upper bound is obtained by exhibiting an estimator having a uniform risk of order $(\ln n/n)^{2/d}$.

Such an estimator (although not computable in practice) was first proposed by Genovese et al. in [21], while another estimator attaining this same minimax rate (computable in practice), and based on the Tangential Delaunay Complex [15], was proposed by Aamari and Levrard [2]. Although being minimax and computable, the Tangential Delaunay Complex depends on the tuning of several parameters (for instance a radius quantifying the size of neighborhoods which are used to compute local PCAs), while those parameters have to be calibrated in a precise manner with respect to the quantities $\tau_{\min}$, $f_{\min}$ and $f_{\max}$ defining the model for the Tangential Delauny Complex to be minimax. However, those quantities are a priori unknown. A first possibility is to estimate those quantities in turn: if procedures are known to estimate the reach (although themselves depending on the tuning of parameters [4, 14]), estimating $f_{\min}$ and $f_{\max}$ appears to be delicate. The problem of the practical choice of the parameters defining the estimator is then raised. This question of the tuning of parameters defining an estimator is not restricted to the framework of manifold estimation, but is a classical problem in statistics.

Let us cite for instance the question of the choice of the bandwidth for kernel density estimation. Let $X_1, \ldots, X_n$ be a n-sample of some law μ having a density f on $\mathbb{R}$, and suppose that we want to recover the value $f(x_0)$ of the density at some fixed point $x_0 \in \mathbb{R}$. A standard method to achieve this goal is to consider the convolution of the empirical measure $\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ by some kernel K_h, where $K_h = h^{-1} K(\cdot/h)$ and K satisfies $\int K = 1$. We then obtain a function $\hat{f} = K_h * \mu_n$. Assume that the density f is of regularity s, that is $f \in \mathcal{C}^s(\mathbb{R})$, the set of $\lfloor s \rfloor$-times differentiable functions, whose $\lfloor s \rfloor$th derivative is $(s - \lfloor s \rfloor)$-Hölder continuous. Then, for a good choice of kernel K, it is optimal to choose the bandwidth h_{opt} of order $c \cdot n^{-1/(2s+1)}$, where c depends on the $\mathcal{C}^s$-norm of f [35, Chapter 1]. The associated risk is then of order $n^{-s/(2s+1)}$, which is the minimax rate of estimation on the class of densities of regularity s. In practice, it is impossible to know exactly the value of s, so that we must find another strategy to choose the bandwidth h. Adaptive methods consist in choosing a bandwidth $\hat{h}$ in a data-dependent way, such that the estimator $\hat{f}_{\hat{h}}$ has a μ-risk almost as good as the optimal estimator $\hat{f}_{h_{\mathrm{opt}}}$ under weak hypotheses on μ. One of such methods, the PCO method (for Penalized Comparison to Overfitting) introduced by Lacour, Massart and Rivoirard [27] consists in comparing each estimator $\hat{f}_h$ to some degenerate estimator $\hat{f}_{h_{\min}}$ for some very small bandwidth $h_{\min}$. The selected bandwidth $\hat{h}$ is chosen among a family $\mathcal{H}$ of bandwidths (all larger than $h_{\min}$), by minimizing a criterion depending on the distance $\|\hat{f}_h - \hat{f}_{h_{\min}}\|_{L_2(\mathbb{R})}$, while penalizing small values of h. Lacour, Massart and Rivoirard then show an oracle inequality for their estimator, that is an inequality of the form

$$\mathbb{E}\|\hat{f}_{\hat{h}} - f\|^2_{L_2(\mathbb{R})} \leq C \min\{\mathbb{E}\|\hat{f}_h - f\|^2_{L_2(\mathbb{R})} : h \in \mathcal{H}\} + C(n, |\mathcal{H}|) \qquad (5)$$

where $C(n, |\mathcal{H}|)$ is a remainder term negligible in front of the optimal risk. Thus, we obtain that $\hat{f}_{\hat{h}}$ has a risk almost as good as the best estimator $\hat{f}_{h_{\mathrm{opt}}}$, while we never had to estimate the parameters defining the statistical model (that is
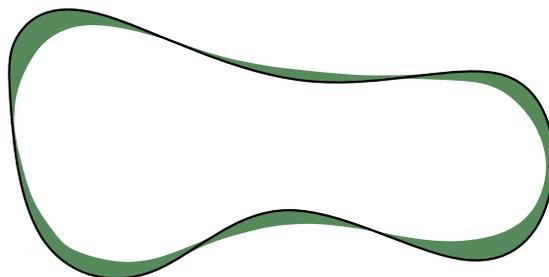
FIG 2. *The t-convex hull* $\mathrm{Conv}(t, A)$ *(in green) of a curve A (in black).*

the regularity s of the density and the $\mathcal{C}^s$-norm of f).

Our main goal is to adapt the PCO method to the manifold inference setting. A first step consists in creating a family of estimators $(\hat{M}_t)_{t \geq 0}$ similar to kernel density estimators, but in the context of manifold estimation. This is made possible with t-convex hulls. For $t \geq 0$, the t-convex hull $\mathrm{Conv}(t, A)$ of a set A is an interpolation between the set A ($t = 0$) and its convex hull $\mathrm{Conv}(A)$ ($t = \infty$). It is defined as

$$\mathrm{Conv}(t, A) := \bigcup_{\substack{\sigma \subseteq A \\ r(\sigma) \leq t}} \mathrm{Conv}(\sigma), \tag{6}$$

where $r(\sigma)$ is the radius of the set σ, that is the radius of the smallest enclosing ball of σ. See Figure 2 for an example. We prove in Section 3 that for $A \subseteq M$, the Hausdorff distance between $\mathrm{Conv}(t, A)$ and M can be efficiently controlled for values of t a little larger than the approximation rate $\varepsilon(A) := \sup\{d(x, A) : x \in M\}$ of A. More precisely, for such values of t, Lemma 3.3 states that $d_H(\mathrm{Conv}(t, A), M) \leq t^2/\tau(M)$. Using this control on the t-convex hull enables us to show that the t-convex hull of the sample $\mathcal{X}_n$ is a minimax estimator on the model $\mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}}$ for a certain choice of t.

Theorem 1.1. *Let α_d be the volume of the d-dimensional unit ball. For the choice of scale $t_n = \frac{7}{4}(3 \ln n / (\alpha_d f_{\min} n))^{1/d}$, we have (for n large enough)*

$$R_n(\mathrm{Conv}(t_n, \mathcal{X}_n), \mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}}) \leq \frac{c_0}{\tau_{\min}(\alpha_d f_{\min})^{2/d}} \left(\frac{\ln n}{n}\right)^{2/d} \tag{7}$$

for some absolute constant c_0. In other words, $\mathrm{Conv}(t_n, \mathcal{X}_n)$ is a minimax estimator of M on $\mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}}$.

To create an adaptive estimator, the next step is to build a selection procedure for the parameter t. An analog of the degenerate estimator $\hat{f}_{h_{\min}}$ is given by the choice $t = 0$, with $\mathrm{Conv}(0, \mathcal{X}_n) = \mathcal{X}_n$. The PCO method therefore suggests comparing the estimators $\mathrm{Conv}(t, \mathcal{X}_n)$ with $\mathcal{X}_n$, that is to study the function $t \mapsto h(t, \mathcal{X}_n) := d_H(\mathrm{Conv}(t, \mathcal{X}_n), \mathcal{X}_n)$. The function $h(\cdot, \mathcal{X}_n)$ was

actually already introduced under the name of "convexity defect function of the set $\mathcal{X}_n$" in a paper by Attali, Lieutier and Salinas [10], where it was used to study the homotopy types of Rips complexes. The convexity defect function is nonnegative, nondecreasing, and satisfies $0 \leq h(t, A) \leq t$ for any set A. For $A = \mathcal{X}_n$, this function is piecewise constant, while it may only change values at $t \in \mathrm{Rad}(\mathcal{X}_n) := \{r(\sigma) : \sigma \subseteq \mathcal{X}_n\}$. We show that the convexity defect function $h(t, \mathcal{X}_n)$ of $\mathcal{X}_n$ at scale t exhibits different behaviors in two different regimes: for $t \leq \varepsilon(\mathcal{X}_n)$ it has a globally linear behavior (that is it stays close to its maximal value t), whereas roughly after $\varepsilon(\mathcal{X}_n)$, it is almost constant. The convexity defect function can be computed using only the dataset, so that we may in practice observe those two regimes. In practice, we fix a value $0 < \lambda < 1$, and let

$$t_\lambda(\mathcal{X}_n) := \inf\{t \in \mathrm{Rad}(\mathcal{X}_n) : h(t, \mathcal{X}_n) \leq \lambda t\}. \tag{8}$$

Our main result states that $t_\lambda(\mathcal{X}_n)$ is a little larger than $\varepsilon(\mathcal{X}_n)$ with high probability, so that we may control the risk of $\hat{M} = \mathrm{Conv}(t_\lambda(\mathcal{X}_n), \mathcal{X}_n)$, without having to know d, $f_{\min}$, $f_{\max}$ or the reach $\tau(M)$, leading to an adaptive estimator in a sense made precise in Theorem 6.2. The estimator $\hat{M}$ is to our knowledge the first minimax adaptive manifold estimator. Our procedure allows us to actually estimate (up to a multiplicative constant arbitrarily close to 1) the approximation rate $\varepsilon(\mathcal{X}_n)$, while scale parameters in computational geometry typically have to be properly tuned with respect to this quantity. The parameter $t_\lambda(\mathcal{X}_n)$ can therefore be used as a hyperparameter in different settings. To illustrate this general idea, we show how to create a data-driven minimax estimator of the tangent spaces of a manifold (see Corollary 6.5).

Related work

"Localized" versions of convex hulls such as the t-convex hulls have already been introduced in the support estimation literature. For instance, slightly modified versions of the t-convex hull have been used as estimators in [5] under the assumption that the support has a smooth boundary and in [33] under reach constraints on the support, with different rates obtained in those models. Selection procedures were not designed in those two papers, and whether our selection procedure leads to an adaptive estimator in those frameworks is an interesting question. The statistical models we study in this article were introduced in [21] and [2], in which manifold estimators were also proposed. If the estimator in [21] is of purely theoretical interest, the estimator proposed by Aamari and Levrard in [2], based on the Tangential Delaunay complex, is computable with $\mathcal{O}(nD2^{\mathcal{O}(d^2)})$ operations. Furthermore, it is a simplicial complex which is known to be ambient isotopic to the underlying manifold M with high probability. It however requires the tuning of several hyperparameters in order to be minimax, which may make its use delicate in practice. In contrast, the t-convex hull estimator with parameter $t_\lambda(\mathcal{X}_n)$ is completely data-driven, computable in polynomial time (see Section 7), while keeping the minimax property. However,

unlike in the case of the Tangential Delaunay complex, we have no guarantees on the homotopy type of the corresponding estimator.

2. Background on submanifold with positive reach

Let us first introduce some notation. The Euclidean norm in $\mathbb{R}^D$ is denoted by $|\cdot|$ and $\langle \cdot, \cdot \rangle$ stands for the dot product. If $A \subseteq \mathbb{R}^D$ and $x \in \mathbb{R}^D$, then $d(x, A) := \inf\{|x - y| : y \in A\}$ is the distance to a set A while $\mathrm{diam}(A) := \sup\{|x - y| : x, y \in A\}$ is its diameter. Given $r \geq 0$, $\mathcal{B}(x, r)$ is the open ball of radius r centered at x and we write $\mathcal{B}_A(x, r)$ for $\mathcal{B}(x, r) \cap A$. We let $\mathcal{M}^d$ be the set of $\mathcal{C}^2$ compact connected d-dimensional submanifolds of $\mathbb{R}^D$ without boundary. If $M \in \mathcal{M}^d$ and $x \in M$, then $T_x M$ is the tangent space of M at x. It is identified with a d-dimensional subspace of $\mathbb{R}^D$, and we write π_x for the orthogonal projection on $T_x M$, while $\pi_x^\perp = \mathrm{id} - \pi_x$ is the projection on the normal space $T_x M^\perp$. The asymmetric Hausdorff distance between sets $A, B \subseteq \mathbb{R}^D$ is defined as $d_H(A|B) := \sup\{d(x, B) : x \in A\}$, while the Hausdorff distance is defined as $d_H(A, B) = \max\{d_H(A|B), d_H(B|A)\}$. For $A \subseteq M$, we denote by $\varepsilon(A) := d_H(A, M)$ the approximation rate of A.

The regularity of a submanifold $M \in \mathcal{M}^d$ is measured by its reach $\tau(M)$. This is the largest number r such that if $d(x, M) < r$ for $x \in \mathbb{R}^D$, then there exists a unique point of M, denoted by $\pi_M(x)$, which is at distance $d(x, M)$ from x. Thus, the projection π_M on the manifold M is well-defined on the r-tubular neigborhood $M^r := \{x \in M : d(x, M) \leq r\}$ for $r < \tau(M)$. The notion of reach was introduced for general sets by Federer in [19], where it is also proven that $\mathcal{C}^2$ compact submanifolds without boundary have positive reach (see [19, p.432]). Different geometric quantities of interest can be bounded in term of the reach. For instance, the volume $\mathrm{Vol}(M)$ of M satisfies

$$\mathrm{Vol}(M) \geq \omega_d \tau(M)^d \tag{9}$$

where ω_d is the volume of a d-dimensional sphere (with equality obtained only for a sphere of radius $\tau(M)$), see [6]. The reach also controls how points on M deviate from their projections on some tangent space.

Lemma 2.1 (Theorem 4.18 in [19]). *For $x, y \in M$, $\left|\pi_x^\perp(y - x)\right| \leq \frac{|y - x|^2}{2\tau(M)}$.*

The following lemma asserts that the projection from a manifold to its tangent space is well-behaved.

Lemma 2.2. *Let $x \in M$.*

1. *Let $y \in \mathbb{R}^D$ with $d(y, M) < \tau(M)$. Then, $\pi_M(y) = x$ if and only if $y - x \in T_x M^\perp$.*
2. *Let $y_1, y_2 \in \mathbb{R}^D$ be two points at distance less than $\gamma < \tau(M)$ from M. Then, $|\pi_M(y_1) - \pi_M(y_2)| \leq \frac{\tau(M)}{\tau(M) - \gamma}|y_1 - y_2|$.*
3. *For $r < \tau(M)/3$, the map $\tilde{\pi}_x : y \mapsto \pi_x(y - x)$ is a diffeomorphism from $\mathcal{B}_M(x, r)$ to its image, and, if $r \leq \tau(M)/2$, we have $\mathcal{B}_{T_x M}(0, 7r/8) \subseteq$*

$\tilde{\pi}_x \left(\mathcal{B}_M(x, r) \right)$. *In particular, if $y \in \mathcal{B}_M(x, 7\tau(M)/24)$, then*

$$\frac{7}{8}|y - x| \leq |\pi_x(y - x)| \leq |y - x|. \tag{10}$$

Proof. • For 1 and 2, see [19, Theorem 4.8].

• We first show that $\tilde{\pi}_x$ is injective on $\mathcal{B}_M(x, \tau(M)/3)$. Assume that $\tilde{\pi}_x(y) = \tilde{\pi}_x(y')$ for some $y \neq y' \in M$. Consider without loss of generality that $|x - y| \geq |x - y'|$. The goal is to show that $|x - y| \geq \tau(M)/3$. If $|x - y| > \tau(M)/2$, the conclusion obviously holds, so we may assume that $|x - y| \leq \tau(M)/2$. Define the angle between T_xM and T_yM as $\|\pi_x - \pi_y\|_{op}$ (where $\| \cdot \|_{op}$ denotes the operator norm). Lemma 3.4 in [12] states that if $|x - y| \leq \tau(M)/2$, then $\angle(T_xM, T_yM) \leq 2\frac{|x-y|}{\tau(M)}$. Also, by definition,

$$\begin{aligned}
\angle(T_xM, T_yM) &\geq \frac{|(\pi_x - \pi_y)(y - y')|}{|y - y'|} \\
&= \frac{|\pi_y(y - y')|}{|y - y'|} \geq \frac{|y - y'| - |\pi_y^{\perp}(y - y')|}{|y - y'|} \\
&\geq 1 - \frac{|y - y'|}{2\tau(M)} \text{ by Lemma 2.1} \\
&\geq 1 - \frac{|x - y|}{\tau(M)} \text{ by the triangle inequality.}
\end{aligned}$$

Therefore, we have $3|x - y|/\tau(M) \geq 1$, i.e. $|x - y| \geq \tau(M)/3$ and $\tilde{\pi}_x$ is injective on $\mathcal{B}_M(x, \tau(M)/3)$. To conclude that $\tilde{\pi}_x$ is a diffeomorphism, it suffices to show that its differential is always invertible. As $\tilde{\pi}_x$ is an affine application, the differential $d\tilde{\pi}_x(y)$ is equal to π_x. Therefore, the Jacobian of the function $\tilde{\pi}_x : M \to T_xM$ at y is given by the determinant of the projection π_x restricted to T_yM. In particular, it is larger than the smallest singular value of $\pi_x \circ \pi_y$ to the power d, which is larger than

$$(1 - \angle(T_xM, T_yM))^d \geq \left(1 - 2\frac{|x - y|}{\tau(M)}\right)^d \geq \left(\frac{1}{3}\right)^d$$

thanks to [12, Lemma 3.4] and using that $|x - y| \leq \tau(M)/3$. In particular, the Jacobian is positive, and $\tilde{\pi}_x$ is a diffeormorphism from $\mathcal{B}_M(x, \tau(M)/3)$ to its image. The second statement is stated in [3, Lemma A.2].

The second inequality of the last statement follows from the projection being 1-Lipschitz continuous. For the first one, let $y \in \mathcal{B}_M(x, 7\tau(M)/24)$, and let $u = \pi_x(y - x)$. The point u is in $\mathcal{B}_{T_xM}(0, h)$ for $h > |u|$. We have $\mathcal{B}_{T_xM}(0, h) \subseteq \tilde{\pi}_x(\mathcal{B}_M(x, 8h/7)) \subseteq \tilde{\pi}_x(\mathcal{B}_M(x, \tau(M)/3))$. As $\tilde{\pi}_x$ is injective on $\mathcal{B}_M(x, \tau(M)/3)$, this means that we necessarily have $y \in \mathcal{B}_M(x, 8h/7)$. Therefore, $|x - y| < 8h/7$, and the conclusion holds by letting h goes to $|u|$. □

It will also be necessary to have precise bounds on the volume of balls on M. As expected, the volume of a small ball is asymptotically equivalent to the volume of an Euclidean ball. Let α_d be the volume of the d-dimensional unit ball.

Lemma 2.3. *Let $r \leq \tau(M)/4$ and $x \in M$. Then,*

$$\left(\frac{47}{48}\right)^d \leq \left(1 - \frac{r^2}{3\tau(M)^2}\right)^d \leq \frac{\mathrm{Vol}\left(\mathcal{B}_M(x,r)\right)}{\alpha_d r^d} \leq \left(1 + \frac{4r^2}{3\tau(M)^2}\right)^d \leq \left(\frac{13}{12}\right)^d.$$
(11)

Proof. The proof of Proposition 8.7 in [2] implies that, if $\tilde{\mathcal{B}}_M(x,r)$ is the geodesic ball centered at x of radius r, then

$$\left(1 - \frac{r^2}{3\tau(M)^2}\right)^d \leq \frac{\mathrm{Vol}\left(\tilde{\mathcal{B}}_M(x,r)\right)}{\alpha_d r^d} \leq \left(1 + \frac{r^2}{\tau(M)^2}\right)^d.$$

As $\tilde{\mathcal{B}}_M(x,r) \subseteq \mathcal{B}_M(x,r)$, we have in particular $\frac{\mathrm{Vol}(\mathcal{B}_M(x,r))}{\alpha_d r^d} \geq \left(1 - \frac{r^2}{3\tau(M)^2}\right)^d$. Furthermore, by [8, Lemma 3.12] and [31, Proposition 6.3], if $|x-y| \leq \tau(M)/4$, then the geodesic distance between x and y is smaller than

$$|x-y|\left(1 + \frac{\pi^2}{50\tau(M)^2}|x-y|^2\right) \leq 1.05|x-y|.$$

This implies that $\mathcal{B}_M(x,r) \subseteq \tilde{\mathcal{B}}_M\left(x, r\left(1 + \frac{\pi^2 r^2}{50\tau(M)^2}\right)\right)$. Therefore,

$$\frac{\mathrm{Vol}\left(\mathcal{B}_M(x,r)\right)}{\alpha_d r^d} \leq \left(\left(1 + \frac{\pi^2 r^2}{50\tau(M)^2}\right)\left(1 + \frac{(1.05r)^2}{\tau(M)^2}\right)\right)^d$$

$$\leq \left(1 + \left(\frac{\pi^2}{50} + (1.05)^2 + \frac{\pi^2(1.05)^2 r^2}{50\tau(M)^2}\right)\frac{r^2}{\tau(M)^2}\right) \leq \left(1 + \frac{4r^2}{3\tau(M)^2}\right)^d,$$

where we used at the last line that $r \leq \tau(M)/4$. $\qquad\square$

3. Approximation of manifolds with t-convex hulls

Let $A \subseteq M$ be a finite set. We investigate in this section how the t-convex hull of A approximates M for different values of t, first in a deterministic setting, then in a random setting. The quantity of interest $d_H(\mathrm{Conv}(t,A), M)$ is by definition the maximum of the two quantities $d_H(\mathrm{Conv}(t,A)|M)$ and $d_H(M|\mathrm{Conv}(t,A))$. The first quantity $d_H(\mathrm{Conv}(t,A)|M)$ is given by the maximum of the distances $d_H(\mathrm{Conv}(\sigma)|M)$ over the simplexes $\sigma \subseteq A$ satisfying $r(\sigma) \leq t$. A naive attempt to bound this quantity leads to a control of order t.

Lemma 3.1. *Let $\sigma \subseteq \mathbb{R}^D$ be a closed set. Then, $d_H(\mathrm{Conv}(\sigma)|\sigma) \leq r(\sigma)$.*

Proof. Let $y \in \mathrm{Conv}(\sigma)$ and let z be the center of the smallest enclosing ball of σ. The half-space $\left\{x \in \mathbb{R}^D : |x-z|^2 - r(\sigma)^2 \leq |x-y|^2 - d(y,\sigma)^2\right\}$ contains σ. It thus contains $\mathrm{Conv}(\sigma)$, and in particular y. Therefore, $d(y,\sigma)^2 \leq r(\sigma)^2 - |y-z|^2 \leq r(\sigma)^2$, concluding the proof. $\qquad\square$

As $\sigma \subseteq M$, we have in particular that $d_H(\text{Conv}(t, A)|M) \leq t$. We can actually obtain a much better bound by exploiting that σ lies on M, which looks locally like a flat space. Consider for instance the case where $\sigma = \{x_0, x_1\}$ is made of two points. Then, the line (x_0, x_1) should be approximately parallel to the tangent space $T_{x_0} M$, with the distance from x_1 to $T_{x_0} M$ being of order $|x_0 - x_1|^2$. As a consequence, the distance from any point of the segment $[x_0, x_1]$ to M is also of order $|x_0 - x_1|^2$. More generally, we have the following result.

Lemma 3.2. *Let* $\sigma \subseteq M$ *with* $r(\sigma) < \tau(M)$ *and let* $y \in \text{Conv}(\sigma)$. *Then,*

$$d(y, M) \leq \tau(M) \left(1 - \sqrt{1 - \frac{r(\sigma)^2}{\tau(M)^2}} \right) \leq \frac{r(\sigma)^2}{2\tau(M)} \left(1 + \frac{r(\sigma)^2}{\tau(M)^2} \right). \qquad (12)$$

In particular, for any $t \geq 0$ *and* $A \subseteq M$,

$$d_H(\text{Conv}(t, A)|M) \leq \frac{t^2}{2\tau(M)} \left[\left(1 + \frac{t^2}{\tau(M)^2} \right) \wedge 2 \right] \leq \frac{t^2}{\tau(M)}. \qquad (13)$$

Proof. Lemma 12 in [10] states that if $\sigma \subseteq M$ satisfies $r(\sigma) < \tau(M)$ and $y \in \text{Conv}(\sigma)$, then,

$$d(y, M) \leq \tau(M) \left(1 - \sqrt{1 - \frac{r(\sigma)^2}{\tau(M)^2}} \right).$$

As $\sqrt{1 - u} \geq 1 - u/2 - u^2/2$ for $u \in [0, 1]$, one obtains the conclusion. $\qquad \square$

The other asymmetric distance $d_H(M | \text{Conv}(t, A))$ is apparently more delicate to handle. It can actually be controlled efficiently if the parameter t is large enough. Indeed, assume that t is large enough so that every point x of M is the projection of some point y of $\text{Conv}(t, A)$. Then we have

$$d(x, \text{Conv}(t, A)) \leq |x - y| = |\pi_M(y) - y| = d(y, M)$$
$$\leq d_H(\text{Conv}(t, A)|M) \leq \frac{t^2}{\tau(M)}. \qquad (14)$$

This suggests defining the parameter

$$t^*(A) := \inf \{ t < \tau(M) : \pi_M(\text{Conv}(t, A)) = M \}. \qquad (15)$$

Lemma 3.2 and (14) imply directly the following lemma.

Lemma 3.3. *Let* $A \subseteq M$ *and* $t > t^*(A)$. *Then,*

$$d_H(\text{Conv}(t, A), M) \leq \frac{t^2}{\tau(M)}. \qquad (16)$$
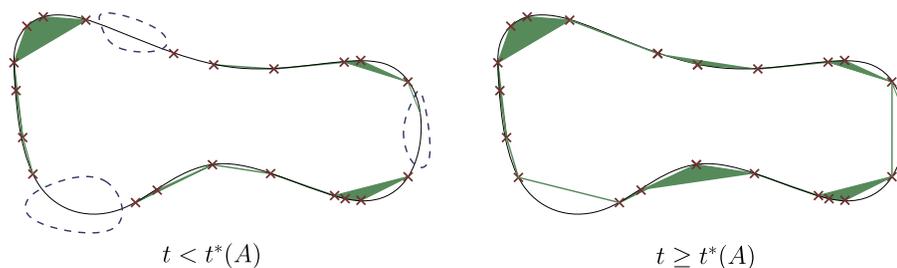
$t < t^*(A)$ $t \geq t^*(A)$

FIG 3. *The t-convex hull of the finite set A (red crosses) is displayed (in green) for two values of t. The black curve represents the (one dimensional) manifold M. On the first display, the value of t is smaller than $t^*(A)$, as there are regions of the manifold (circled in blue) which are not attained by the projection π_M restricted to the t-convex hull. The value of t is larger than $t^*(A)$ on the second display.*

A crucial result in the analysis of the t-convex hull estimator is given by the next proposition, that indicates that the quantity $t^*(A)$ is almost equal to the approximation rate $\varepsilon(A)$.

Proposition 3.4. *Let $A \subseteq M$ be a finite set. Then, $\varepsilon(A) \leq t^*(A)\left(1 + \frac{t^*(A)}{\tau(M)}\right)$. Furthermore, if $\varepsilon(A) < \tau(M)/8$, then, $t^*(A) \leq \varepsilon(A)\left(1 + 6\frac{\varepsilon(A)}{\tau(M)}\right)$*

The proof of Proposition 3.4 relies on considering Delaunay triangulations. Given $d+1$ points σ in $\mathbb{R}^d$ that do not lie on a hyperplane, there exists a unique ball that contains the points on its boundary. It is called the circumball of σ, and its radius is called the circumradius $\mathrm{circ}(\sigma)$ of σ. Given a finite set $A \subseteq \mathbb{R}^d$ that does not lie on a hyperplane, there exists a triangulation of A, called the Delaunay triangulation, such that for each simplex σ in the triangulation, the circumball of σ contains no point of A in its interior. Note that there may exist several Delaunay triangulations of a set A, should the set A not be in general position. With a slight abuse, we will still refer to "the" Delaunay triangulation of A, by simply choosing a Delaunay triangulation among the possible ones should several exist. If the set A lies on a lower dimensional subspace, we consider the Delaunay triangulation of A in the affine vector space spanned by A. Therefore, for every set A, the Delaunay triangulation is well defined (for instance, the Delaunay triangulation of three points aligned in the plane is the 1-dimensional triangulation obtained by joining the middle point with the two others).

Proof. Let $x \in M$ be such that $d(x,A) = \varepsilon(A)$. By definition, there exists a simplex $\sigma \subseteq A$ of radius smaller than $t^*(A)$ with $x = \pi_M(y)$ for some point $y \in \mathrm{Conv}(\sigma)$. We have, using Lemma 3.1 and Lemma 3.2,

$$\varepsilon(A) = d(x,A) \leq |x - y| + d(y,A) \leq \frac{t^*(A)^2}{\tau(M)} + t^*(A)$$

proving the first inequality.

To prove the other inequality, without loss of generality, we assume that $0 \in M$ and we show that $0 \in \pi_M(\mathrm{Conv}(t, A))$ for $t = \varepsilon(A)(1 + 6\varepsilon(A)/\tau(M))$. Let $\tilde{A} = \pi_0(A \cap \mathcal{B}(0, R))$ for $R = \varepsilon(A)\left(2 + c_0\varepsilon(A)/\tau(M)\right)$ and $c_0 = 32/49$. Note that the condition $\varepsilon(A) \leq \tau(M)/8$ implies that $R < 7\tau(M)/24$. We first state two lemmas.

Lemma 3.5. *Assume that* $\varepsilon(A) \leq 7\tau(M)/24$. *Let* $\tilde{x} \in T_0M$ *with* $|\tilde{x}| \leq \varepsilon(A)$. *Then* $d(\tilde{x}, \tilde{A}) \leq \varepsilon(A)$.

Proof. By continuity, it suffices to prove the claim for $|\tilde{x}| < \varepsilon(A)$. In this case, according to Lemma 2.2, if $\varepsilon(A) \leq 7\tau(M)/24$, then there exists $x \in \mathcal{B}_M(0, 8\varepsilon(A)/7)$ with $\pi_0(x) = \tilde{x}$. Furthermore, by Lemma 2.1,

$$|x| \leq |\tilde{x}| + |x - \tilde{x}| \leq \varepsilon(A) + \frac{|x|^2}{2\tau(M)} \leq \varepsilon(A)\left(1 + \frac{32\varepsilon(A)}{49\tau(M)}\right).$$

We have $d(x, A) = |x - a|$ for some point $a \in A$, and $|a| \leq |x - a| + |x| \leq \varepsilon(A)\left(2 + c_0\varepsilon(A)/\tau(M)\right)$. As $\pi_0(a) \in \tilde{A}$, we have $d(\tilde{x}, \tilde{A}) \leq |\tilde{x} - \pi_0(a)| \leq |x - a| = d(x, A) \leq \varepsilon(A)$. $\qquad \square$

Lemma 3.6. *Let* $V \subseteq \mathbb{R}^d$ *be a finite set and* $t > 0$. *If* $d_H(\mathcal{B}(0, t)|V) \leq t$, *then* $0 \in \mathrm{Conv}(V)$.

Proof. We prove the contrapositive. If $0 \notin \mathrm{Conv}(V)$, then there exists an open half-space which contains V. Let x be the unit vector orthogonal to this half-space. Then, $d(tx, V) > t$. $\qquad \square$

Apply Lemma 3.6 to $V = \tilde{A}$ and $t = \varepsilon(A)$. For $\tilde{x} \in \mathcal{B}_{T_0M}(0, \varepsilon(A))$, we have $d(\tilde{x}, \tilde{A}) \leq \varepsilon(A)$ according to Lemma 3.5. Therefore, we have $0 \in \mathrm{Conv}(\tilde{A})$. Consider the Delaunay triangulation of $\tilde{A}$. The point 0 belongs to the convex hull of some simplex $\tilde{\sigma}$ of the triangulation, with circumradius $\mathrm{circ}(\tilde{\sigma})$ and center of the circumball $\tilde{q}$. The simplex $\tilde{\sigma}$ corresponds to some simplex σ in A, and the point 0 is equal to $\pi_0(y)$ for some point $y \in \mathrm{Conv}(\sigma)$. By Lemma 2.2, we actually have $\pi_M(y) = 0$, and to conclude, it suffices to show that $r(\sigma) \leq \varepsilon(A)\left(1 + 6\frac{\varepsilon(A)}{\tau(M)}\right)$. To do so, we use the next lemma (recall that $\sigma \subseteq \mathcal{B}_M(0, R)$ with $R < 7\tau(M)/24$).

Lemma 3.7. *Let* $\sigma \subseteq \mathcal{B}_M(0, 7\tau(M)/24)$ *and* $\tilde{\sigma} = \tilde{\pi}_0(\sigma)$. *Assume that* $0 \in \mathrm{Conv}(\tilde{\sigma})$. *Then,*

$$r(\tilde{\sigma}) \leq r(\sigma) \leq r(\tilde{\sigma})\left(1 + 6\frac{r(\tilde{\sigma})}{\tau(M)}\right). \tag{17}$$

Proof. As the projection is 1-Lipschitz, it is clear that $r(\tilde{\sigma}) \leq r(\sigma)$. Let us prove the other inequality. Let $\sigma = \{y_0, \ldots, y_k\}, \tilde{\sigma} = \{\tilde{y}_0, \ldots, \tilde{y}_k\}$ and fix $0 \leq i \leq k$. As $y_i \in \mathcal{B}_M(0, 7\tau(M)/24)$, we have by (10)

$$|y_i| \leq \frac{8}{7}|\tilde{y}_i| \leq \frac{16}{7}r(\tilde{\sigma}). \tag{18}$$
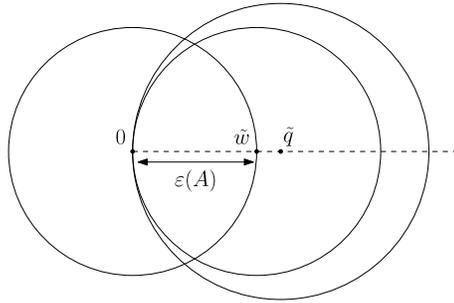
FIG 4. *If $|\tilde{q}| > \varepsilon(A)$, then the ball $\mathcal{B}_{T_0 M}(\tilde{q}, |\tilde{q}|)$ contains a ball of radius $\varepsilon(A)$ centered at a point (here denoted by $\tilde{w}$) at distance less than $\varepsilon(A)$ from 0.*

where we used that $|\tilde{y}_i| \leq 2r(\tilde{\sigma})$ as $0 \in \mathrm{Conv}(\tilde{\sigma})$. Let $\tilde{z}$ be the center of the minimum enclosing ball of $\tilde{\sigma}$. Write $\tilde{z} = \sum_{j=0}^{k} \lambda_j \tilde{y}_j$ as a convex combination of the $\tilde{y}_j$s and let $z = \sum_{j=0}^{k} \lambda_j y_j \in \mathrm{Conv}(\sigma)$. Then, we have

$$
\begin{aligned}
|z - y_i| &\leq |z - \tilde{z}| + |\tilde{z} - \tilde{y}_i| + |\tilde{y}_i - y_i| \\
&\leq \sum_{j=0}^{k} \lambda_j \, |y_j - \tilde{y}_j| + r(\tilde{\sigma}) + \frac{|y_i|^2}{2\tau(M)} \text{ using Lemma 2.1} \\
&\leq \sum_{j=0}^{k} \lambda_j \frac{|y_j|^2}{2\tau(M)} + r(\tilde{\sigma}) + \frac{128}{49} \frac{r(\tilde{\sigma})^2}{\tau(M)} \text{ using Lemma 2.1 and (18)} \\
&\leq r(\tilde{\sigma}) + \frac{256}{49} \frac{r(\tilde{\sigma})^2}{\tau(M)} \leq r(\tilde{\sigma}) + 6 \frac{r(\tilde{\sigma})^2}{\tau(M)} \text{ using (18).}
\end{aligned}
$$

We obtain the conclusion as σ is included in the ball of radius $\max_i |z - y_i|$ and center z. □

Using the previous lemma, we are left with showing that $r(\tilde{\sigma}) \leq \varepsilon(A)$. We will actually show the stronger inequality $\mathrm{circ}(\tilde{\sigma}) \leq \varepsilon(A)$ (the radius of a set is always smaller than its circumradius). As 0 is in the circumball (that is centered at $\tilde{q}$), the ball centered at $\tilde{q}$ of radius $|\tilde{q}|$ does not intersect $\tilde{A}$. This enforces $|\tilde{q}| \leq \varepsilon(A)$: otherwise, there would exist a ball not intersecting $\tilde{A}$, of radius $\varepsilon(A)$, and whose center is at distance less than $\varepsilon(A)$ from 0, a contradiction with Lemma 3.5 (see Figure 4). As $|\tilde{q}| \leq \varepsilon(A)$, we obtain, once again according to Lemma 3.5, that $\mathrm{circ}(\tilde{\sigma}) = d(\tilde{q}, \tilde{A}) \leq \varepsilon(A)$ concluding the proof. □

Remark 3.8. *In the case where the dimension d is known, one can consider a variant of the t-convex hull, $\mathrm{Conv}_d(t, A)$, where one restricts the union to be over simplices of dimension less than d. The set $\mathrm{Conv}_d(t, A)$ is simpler to compute as it contains less simplices (see Section 7). Furthermore, if $t_d^*(A) := \inf \{t : \pi_M(\mathrm{Conv}_d(t, A)) = M\}$, then both Lemma 3.3 and Proposition*

3.4 hold with $t_d^*(A)$ and $\mathrm{Conv}_d(t, A)$ instead of $t^*(A)$ and $\mathrm{Conv}(t, A)$. Indeed, only simplices of dimension less than d (corresponding to simplices of a Delaunay triangulation on a tangent space) were considered in the previous proof.

We have now shown that the quality of the t-convex hull on A can be controlled for $t \geq \varepsilon(A)(1 + 6\varepsilon(A)/\tau(M))$ (that is slightly larger than the approximation rate $\varepsilon(A)$). In a random setting, the approximation rate is known to be of order $(\ln n/n)^{1/d}$: this is enough to show that the t-convex hull is a minimax estimator. Recall the definition of the statistical model $\mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}}$ from the introduction: it consists of laws μ supported on some manifold $M \in \mathcal{M}^d$ with $\tau(M) \geq \tau_{\min}$, having a density f lower bounded by $f_{\min}$ and upper bounded by $f_{\max}$. The minimax result will actually hold on the larger model $\mathcal{Q}^d_{\tau_{\min}, f_{\min}} := \bigcup_{f_{\max}} \mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}}$ (that is without imposing any upper bound on f).

Let $\mu \in \mathcal{Q}^d_{\tau_{\min}, f_{\min}}$ and let $\mathcal{X}_n$ be a n-sample from law μ. We consider the estimator $\mathrm{Conv}(t, \mathcal{X}_n)$. Note first that $\mathrm{Conv}(t, \mathcal{X}_n)$ is indeed an estimator, that is the application

$$(x_1, \ldots, x_n) \in \left(\mathbb{R}^D\right)^n \mapsto \mathrm{Conv}(t, \{x_1, \ldots, x_n\})$$

is measurable (with respect to the Borel σ-field associated with the metric d_H on the set $\mathcal{K}\left(\mathbb{R}^D\right)$ of all nonempty compact subsets of $\mathbb{R}^D$). Indeed, for E a measurable subset of $\mathcal{K}\left(\mathbb{R}^D\right)$ and $A, B \in \mathcal{K}\left(\mathbb{R}^D\right)$, introduce the notation $G_E(A, B) = A$ if $A \in E$ and B otherwise. This function is measurable, and $\mathrm{Conv}(t, \{x_1, \ldots, x_n\})$ can be written as

$$\bigcup_{I \subseteq \{1, \ldots, n\}} G_E\left(\mathrm{Conv}(\{x_i\}_{i \in I}), \{x_i\}_{i \in I}\right)$$

where E is the subset of $\mathcal{K}\left(\mathbb{R}^D\right)$ given by $\left\{K \in \mathcal{K}\left(\mathbb{R}^D\right) : r(K) \leq t\right\}$, which is closed [10, Lemma 16]. As the functions $\cup$ and Conv are measurable, the measurability follows [1, Proposition III.7].

For a fixed $t > 0$, we obtain the following control of $\mathbb{E}\left[d_H(\mathrm{Conv}(t, \mathcal{X}_n), M)\right]$.

$$
\begin{aligned}
\mathbb{E}\left[d_H(\mathrm{Conv}(t, \mathcal{X}_n), M)\right] &= \mathbb{E}\left[d_H(\mathrm{Conv}(t, \mathcal{X}_n), M)\mathbf{1}\left\{t \geq t^*(\mathcal{X}_n)\right\}\right] \\
&\quad + \mathbb{E}\left[d_H(\mathrm{Conv}(t, \mathcal{X}_n), M)\mathbf{1}\left\{t < t^*(\mathcal{X}_n)\right\}\right] \\
&\leq \frac{t^2}{\tau(M)} + \mathrm{diam}(M)\mathbb{P}\left(t^*(\mathcal{X}_n) > t\right)
\end{aligned}
$$

By Proposition 3.4, if $\varepsilon(\mathcal{X}_n) < \tau(M)/8$, then

$$t^*(\mathcal{X}_n) \leq \varepsilon(\mathcal{X}_n)\left(1 + 6\frac{\varepsilon(\mathcal{X}_n)}{\tau(M)}\right) \leq \frac{7}{4}\varepsilon(\mathcal{X}_n).$$

Therefore, if t is small enough,

$$\mathbb{P}\left(t^*(\mathcal{X}_n) > t\right) \leq \mathbb{P}\left(\varepsilon(\mathcal{X}_n) > \tau(M)/8\right) + \mathbb{P}\left(\varepsilon(\mathcal{X}_n) > 4t/7\right) \leq 2\mathbb{P}\left(\varepsilon(\mathcal{X}_n) > 4t/7\right).$$

We obtain

$$\mathbb{E}\left[d_H(\mathrm{Conv}(t, \mathcal{X}_n), M)\right] \leq \frac{t^2}{\tau(M)} + 2\,\mathrm{diam}(M)\mathbb{P}\left(\varepsilon(\mathcal{X}_n) > 4t/7\right) \qquad (19)$$

Hence, to control the risk, it suffices to bound the tail of $\varepsilon(\mathcal{X}_n)$.

Proposition 3.9. *Let* $\mu \in \mathcal{Q}^d_{\tau_{\min}, f_{\min}}$ *and let* $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ *be a* n-*sample of law* μ. *If* $r \leq \tau_{\min}/4$, *then, for any* $\eta \in (0, 1)$

$$\mathbb{P}\left(\varepsilon(\mathcal{X}_n) > r\right) \leq \frac{c_{d,\eta}}{f_{\min}r^d} \exp\left(-n\alpha_d f_{\min}\left(1 - \frac{r^2}{3\tau_{\min}^2}\right)^d \eta r^d\right), \qquad (20)$$

where $c_{d,\eta}$ *depends on* d *and* η. *Furthermore, for any* $a > 0$, *for* n *large enough (with respect to* d, $f_{\min}, \tau_{\min}$ *and* a), *with probability* $1 - c(\ln n)^{d-1}n^{1-a}$ *(where* c *depends also on those parameters), we have*

$$\varepsilon(\mathcal{X}_n) \leq \left(\frac{a \ln n}{\alpha_d f_{\min}n}\right)^{1/d}. \qquad (21)$$

Proof. A measure ν is said to be (a, b)-standard at scale r_0 if $\nu(\mathcal{B}(x, r)) \geq ar^b$ for all $r \leq r_0$ and x in the support of ν. Let $\mu \in \mathcal{Q}^d_{\tau_{\min}, f_{\min}}$ with support M. Lemma 2.3 indicates that the measure μ is (a, b)-standard at scale r_0 for any $r_0 \leq \tau(M)/4$, with $a = f_{\min}\alpha_d\left(1 - \frac{r_0^2}{3\tau(M)^2}\right)^d$ and $b = d$. It is stated in the proof [1, Proposition III.14] that for such a measure, and for any $\delta \leq 2r_0$ with $0 < r - \delta \leq r_0$, we have

$$\mathbb{P}\left(\varepsilon(\mathcal{X}_n) > r\right) \leq \frac{2^b}{a\delta^b} \exp\left(-na(r - \delta)^b\right).$$

Letting $r = r_0$ and $\delta = \left(1 - \eta^{1/d}\right)r$ for some $\eta \in (0, 1)$, we obtain that

$$\mathbb{P}\left(\varepsilon(\mathcal{X}_n) > r\right)$$
$$\leq \frac{\left(2/\left(1 - \eta^{1/d}\right)\right)^d}{f_{\min}\alpha_d\left(1 - \frac{r^2}{3\tau(M)^2}\right)^d r^d} \exp\left(-nf_{\min}\alpha_d\left(1 - \frac{r^2}{3\tau(M)^2}\right)^d \eta r^d\right)$$
$$\leq \frac{c_0^d}{f_{\min}\alpha_d r^d} \exp\left(-nf_{\min}\alpha_d\left(1 - \frac{r^2}{3\tau(M)^2}\right)^d \eta r^d\right)$$

$$\hspace{12cm} (22)$$

for $c_0 = 96/\left(47\left(1 - \eta^{1/d}\right)\right)$, where we used at the last line that $r \leq \tau(M)/4$.

To prove the second statement, let $r = \left(\frac{a \ln n}{\alpha_d f_{\min}n}\right)^{1/d}$. Then, $n\alpha_d f_{\min}r^d = a \ln n$. Letting $\eta = 1 - 1/\ln n$, we obtain that

$$nf_{\min}\alpha_d\left(1-\frac{r^2}{3\tau(M)^2}\right)^d\eta r^d=(a\ln n)\left(1-\frac{1}{\ln n}\right)\left(1-c\left(\frac{\ln n}{n}\right)^{2/d}\right)$$
$$\geq(a\ln n)-C_a.$$

In particular, the upper bound in (22) is of order $(\ln n)^{d-1}n^{1-a}$.   □

Choose t such that $4t/7=\left(\frac{3\ln n}{\alpha_d f_{\min n}n}\right)^{1/d}$. Then, according to Proposition 3.9, we have $\mathbb{P}\left(\varepsilon(\mathcal{X}_n)>32t/7\right)\leq c(\ln n)^{d-1}n^{-2}$. As $\mathrm{diam}(M)$ is also bounded by a constant depending on $\tau_{\min},f_{\min}$ and d (see [1, Lemma III.24]), we obtain Theorem 1.1 from (19) (without even the need of assuming that the density f is upper bounded).

4. Selection procedure for the t-convex hulls

Assuming that we have observed a n-sample $\mathcal{X}_n$ having a distribution $\mu\in\mathcal{P}^d_{\tau_{\min},f_{\min}}$, we were able in the previous section to build a minimax estimator of the underlying manifold M. The tuning of this estimator requires the knowledge of d and $f_{\min}$: if the dimension d can be efficiently estimated, this is not the case for $f_{\min}$, which will likely not be accessible in practice. An idea to overcome this issue is to design a selection procedure for the family of estimators $(\mathrm{Conv}(t,\mathcal{X}_n))_{t\geq 0}$. As the loss of the estimator $\mathrm{Conv}(t,\mathcal{X}_n)$ is controlled efficiently for $t\geq t^*(\mathcal{X}_n)$ a good idea is to select a scale t larger than $t^*(\mathcal{X}_n)$. We however do not have access to this quantity based on the observations $\mathcal{X}_n$, as the manifold M is unknown. To select a scale close to $t^*(\mathcal{X}_n)$, we monitor how the estimators $\mathrm{Conv}(t,\mathcal{X}_n)$ deviate from $\mathcal{X}_n$ as t increases. Namely, we use the convexity defect function introduced in [10].

Definition 4.1. *Let $A\subseteq\mathbb{R}^D$ and $t>0$. The convexity defect function at scale t of A is defined as*

$$h(t,A):=d_H(\mathrm{Conv}(t,A),A). \tag{23}$$

As its name indicates, the convexity defect function measures the (lack of) convexity of a set A at a given scale t. The next proposition states preliminary results on the convexity defect function.

Proposition 4.2. *Let $A\subseteq\mathbb{R}^D$ be a closed set and $t\geq 0$.*

1. *We have $0\leq h(t,A)\leq t$.*
2. *The set A is convex if and only if $h(\cdot,A)\equiv 0$.*
3. *If $M\in\mathcal{M}^d$, then $h(t,M)\leq\frac{t^2}{2\tau(M)}\left(1+\frac{t^2}{\tau(M)^2}\right)$.*

Proof. Point 1 follows from Lemma 3.1. Point 2 is clear and Point 3 is a consequence of Lemma 3.2.   □
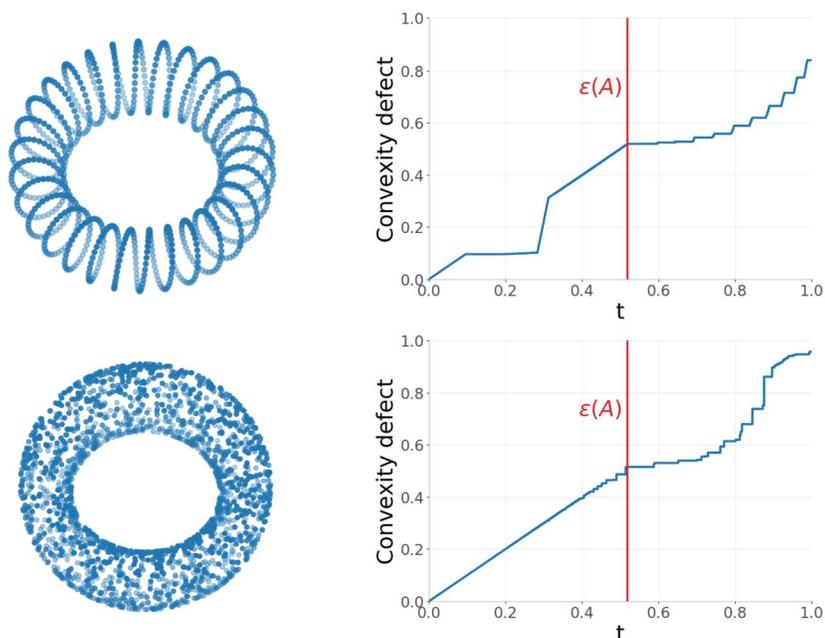
FIG 5. *Two subsets of the torus having the same approximation rate, but whose convexity defect functions exhibit different behaviors on* $[0, t^*(A)]$.

As expected, the convexity defect of a convex set is null, whereas for small values of t, the convexity defect of a manifold $h(t, M)$ is very small (compared to the maximum value possible, which is t): when looked at locally, M is "almost flat" (and thus "almost locally convex"). As already noted in the introduction, if A is a finite set, then the convexity defect function is a piecewise constant function, whose value may only change at t if $t \in \text{Rad}(A) := \{r(\sigma) : \sigma \subseteq A\}$.

For a set $A \subseteq M$, we recover the subquadratic behavior of the convexity defect function for values of t above the threshold value $t^*(A)$. Namely, we have the following proposition.

Proposition 4.3. *Let* $A \subseteq M$. *For* $t^*(A) < t < \tau(M)$,

$$h(t, A) \le \frac{t^2}{2\tau(M)} \left(1 + \frac{t^2}{\tau(M)^2}\right) + t^*(A) \left(1 + \frac{t^*(A)}{\tau(M)}\right). \tag{24}$$

Proof. By using that $h(t, A) \le t$ and Lemma 3.3, for any $t^*(A) < s < t$,

$$
\begin{aligned}
h(t, A) &= d_H(\text{Conv}(t, A), A) \\
&\le d_H(\text{Conv}(t, A), M) + d_H(M, \text{Conv}(s, A)) + d_H(\text{Conv}(s, A), A) \\
&\le \frac{t^2}{2\tau(M)} \left(1 + \frac{t^2}{\tau(M)^2}\right) + \frac{s^2}{2\tau(M)} \left(1 + \frac{s^2}{\tau(M)^2}\right) + s
\end{aligned}
$$

The conclusion is obtained by letting s go to $t^*(A)$.   □

For $0 < t < t^*(A)$, the convexity defect function may exhibit very different behaviors, as shown in Figure 5. However, when the set $A = \mathcal{X}_n$ is a random n-sample, it appears that the graph of the convexity defect function stays close to the diagonal $\{x = y\}$ for small values of t. This is explained by the fact that for two points X_1, X_2 in the sample at very small distance $2t$ from one another, it is very unlikely that there is a third point at distance of order t from X_1 and X_2, so that $d_H(\mathrm{Conv}(\{X_1, X_2\})|\mathcal{X}_n) = d_H(\mathrm{Conv}(\{X_1, X_2\})|\{X_1, X_2\}) = t$.

This suggests the following strategy to select a value of t larger than $t^*(\mathcal{X}_n)$ using the convexity defect function:

Definition 4.4. *Let $A \subseteq M$ be a finite set and $0 < \lambda \leq 1$. We define*

$$t_\lambda(A) := \inf\{t \in \mathrm{Rad}(A) : h(t, A) \leq \lambda t\}. \tag{25}$$

Restricting to values $t \in \mathrm{Rad}(A)$ is necessary, for otherwise we would always have $t_\lambda(A) = 0$ (as $h(t, A) = 0$ for t small enough). Proposition 4.3 implies that $t_\lambda(A)$ cannot be too large. More precisely, we have the following lemma.

Lemma 4.5. *Let $A \subseteq M$ with $t^*(A) \leq \lambda^2 \tau(M)/4$. Let $r_0 = \frac{t^*(A)}{\lambda}\left(1 + \frac{8}{\lambda^2}\frac{t^*(A)}{\tau(M)}\right)$ and $r_1 = \lambda\tau(M)/2$. If $t \in \mathrm{Rad}(A) \cap [r_0, r_1]$, then $t_\lambda(A) \leq t$.*

Proof. By Proposition 4.3, we have, for $t^*(A) < t \leq \lambda\tau(M)/2$,

$$
\begin{aligned}
h(t, A) &\leq \frac{t^2}{2\tau(M)}\left(1 + \frac{t^2}{\tau(M)^2}\right) + t^*(A)\left(1 + \frac{t^*(A)}{\tau(M)}\right) \\
&\leq \frac{t^2}{2\tau(M)}\left(1 + \lambda^2/4\right) + t^*(A)\left(1 + \frac{t^*(A)}{\tau(M)}\right) - \lambda t + \lambda t =: P(t) + \lambda t.
\end{aligned}
$$

Let $u = 2t^*(A)\left(1 + \frac{t^*(A)}{\tau(M)}\right)\left(1 + \lambda^2/4\right)/\left(\lambda^2\tau(M)\right)$. The condition $t^*(A) \leq \lambda^2\tau(M)/4$ ensures that $u \leq 1$. The quantity $P(t)$ is nonpositive if t is between t_0 and t_1, where

$$t_0 = \frac{\tau(M)\lambda}{1 + \lambda^2/4}(1 - \sqrt{1-u}) \quad \text{and} \quad t_1 = \frac{\tau(M)\lambda}{1 + \lambda^2/4}(1 + \sqrt{1-u}).$$

We have $t_1 \geq r_1$ and, using the inequality $\sqrt{1-u} \geq 1 - \frac{u}{2} - \frac{u^2}{2}$ for $0 \leq u \leq 1$, we obtain that $t_0 \leq r_0$. Therefore, any $t \in [r_0, r_1]$ satisfies $h(t, A) \leq \lambda t$ (note that $r_0 > t^*(A)$). In particular, if t is also in $\mathrm{Rad}(A)$, we have $t_\lambda(A) \leq t$.   □

Our main theorem states that, with high probability, the parameter $t_\lambda(\mathcal{X}_n)$ is larger than $t^*(\mathcal{X}_n)$

Theorem 4.6.   *1. Let $\mu \in \mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}}$. Let $0 < b \leq 2$ and let $\mathcal{X}_n$ be a n-sample of law μ. Let $a = (d-1) \vee 2$ if $b = 2$, and $a = d-1$ otherwise. For n large enough, and with probability larger than $1 - c(\ln n)^a n^{-b}$, we have for $0 < \lambda < (1 + b)^{-1/d}$,*

$$t^*(\mathcal{X}_n) \leq t_\lambda(\mathcal{X}_n) \leq \frac{t^*(\mathcal{X}_n)}{\lambda}\left(1 + C\left(\frac{(\ln n)^2}{n}\right)^{1/d}\right) \tag{26}$$

where the constant c depends on b, and μ, and C depends on $f_{\min}$, $f_{\max}$, d, $\tau_{\min}$ and λ.

2. Furthermore, if μ is the uniform distribution on the circle of radius $\tau_{\min}$, then, for $\lambda > (1+b)^{-1}$, we have

$$\mathbb{P}\left(t^*(\mathcal{X}_n) > t_\lambda(\mathcal{X}_n)\right) \geq cn^{-b} \tag{27}$$

for some constant c depending on $\tau_{\min}$ and b.

Inequality (27) implies that the probability $1 - c(\ln n)^a n^{-b}$ appearing in the theorem is close to being tight.

Proof of the upper bound in (26)

Let $\mu \in \mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}}$ be a probability distribution with support M and density f. We assume without loss of generality that $f_{\min}$ is the essential infimum of f. Recall the notation $r_1 = \lambda\tau(M)/2$ and $r_0 = \frac{t^*(\mathcal{X}_n)}{\lambda}\left(1 + \frac{8}{\lambda^2}\frac{t^*(\mathcal{X}_n)}{\tau(M)}\right)$ from Lemma 4.5. The proof of the upper bound is based on the following lemma.

Lemma 4.7. *There exists a positive constant $\beta > 0$ (depending on $f_{\min}, f_{\max}, d$ and $\tau_{\min}$) such that the following holds. Let $\alpha > 0$ and let $I = [a, b]$ be an interval of length at least $\ell = \alpha\left(\frac{\ln n}{n}\right)^{2/d}$ with $b \leq \beta\alpha\left(\frac{\ln n}{n}\right)^{1/d}$ and $a \geq \ell/2$. Then, the probability that $\mathrm{Rad}(\mathcal{X}_n)$ does not intersect I is smaller than $n^{-1/2}$.*

Before proving the lemma, let us use it to obtain the upper bound in (26). By Proposition 3.9 and Proposition 3.4, we have $t^*(\mathcal{X}_n) \leq \left(\frac{4\ln n}{\alpha_d f_{\min} n}\right)^{1/d}$ everywhere but on a set of probability smaller than n^{-2}. We will assume that this condition is satisfied. In particular, the condition $t^*(\mathcal{X}_n) \leq \lambda^2\tau(M)/4$ of Lemma 4.5 is satisfied. Let $u = \delta\left(\frac{(\ln n)^2}{n}\right)^{1/d}$ (for some constant δ to fix) and let

$$R_0 := r_0(1+u) \leq \frac{t^*(\mathcal{X}_n)}{\lambda}\left(1 + 2\delta\left(\frac{(\ln n)^2}{n}\right)^{1/d}\right) \leq \frac{2}{\lambda}\left(\frac{4\ln n}{\alpha_d f_{\min} n}\right)^{1/d} \leq r_1.$$

Lemma 4.8. *Let $A \subseteq M$ be a finite set of cardinality n. Then,*

$$\varepsilon(A) \geq c_d\tau(M)n^{-1/d}.$$

Proof. If $\varepsilon(A) \geq \tau(M)/4$, the conclusion holds. Otherwise, as M is included in $\bigcup_{x \in A} \mathcal{B}_M(x, \varepsilon(A))$, one has $\mathrm{Vol}(M) \leq nc_d\varepsilon(A)^d$ (using Lemma 2.3). We conclude with inequality (9). $\square$

According to Lemma 4.8 and Proposition 3.4, the interval $[r_0, R_0]$ is of length $r_0 u \geq C_1 \delta \left(\frac{\ln n}{n}\right)^{2/d} =: \ell$ for some constant C_1. Choose δ large enough so that $\frac{2}{\lambda} \left(\frac{4}{\alpha_d f_{\min}}\right)^{1/d} \leq \beta C_1 \delta$. Then, as $r_0 \geq \ell/2$ (once again by Lemma 4.8), one can apply Lemma 4.7: the interval $[r_0, R_0]$ intersects $\mathrm{Rad}(\mathcal{X}_n)$ with probability $1 - n^{-2}$. Lemma 4.5 then yields the conclusion.

Proof of Lemma 4.7. Let $I_k = [k\ell/2, (k+1)\ell/2]$ for k an integer. Assume that we show that $\mathrm{Rad}(\mathcal{X}_n)$ intersects every interval I_k for $k = 1, \ldots, K$, where K is chosen so that $b \leq \beta\alpha \left(\frac{\ln n}{n}\right)^{1/d} \leq (K+1)\ell/2$, say $K+1 = \left\lceil 2\beta\alpha \left(\frac{\ln n}{n}\right)^{1/d} / \ell \right\rceil = \left\lceil 2\beta(n/\ln n)^{1/d} \right\rceil$. As the interval I is of length at least ℓ, and as $\ell/2 \leq a$, the interval I contains one of the interval I_k for some $1 \leq k \leq K$. In particular, the interval I also intersects $\mathcal{X}_n$. Therefore, it suffices to bound the probability that $\mathrm{Rad}(\mathcal{X}_n)$ does not intersect I_k. If we show that this probability is of order at most n^{-3}, we may then conclude by a union bound: the probability that $\mathrm{Rad}(\mathcal{X}_n)$ intersects all the I_k is larger than $1 - 2Kn^{-3} \geq 1 - 4\beta n^{1/d-3}/(\ln n)^{1/d} \geq 1 - n^{-2}$.

To bound the probability that $\mathrm{Rad}(\mathcal{X}_n)$ does not intersect I_k, we split the set $\mathcal{X}_n$ into two groups: the set $\mathcal{X}_n^0 = \{X_1, \ldots, X_L\}$ (for some integer L to fix), and the set $\mathcal{X}_n^1 = \{X_{L+1}, \ldots, X_n\}$. If some distance $|X_i - X_j|$ is between $k\ell$ and $(k+1)\ell$, then $\mathrm{Rad}(\mathcal{X}_n)$ intersects I_k. We will show that it is very likely that $|X_i - X_j| \in [k\ell, (k+1)\ell]$ for some $i \leq L$ and $j > L$. To do so, we consider the ball B_i centered at the point X_i, of radius $(k+1)\ell$. Let Y be a point sampled according to μ, conditioned on being in B_i. Then, according to Lemma 2.3, we have

$$\mathbb{P}\left(|Y - X_i| \in [k\ell, (k+1)\ell] | X_i\right) = \frac{\mu\left(\mathcal{B}\left(X_i, (k+1)\ell\right) \setminus \mathcal{B}\left(X_i, k\ell\right)\right)}{\mu\left(\mathcal{B}\left(X_i, (k+1)\ell\right)\right)}$$

$$\geq \frac{c_d f_{\min}}{f_{\max}(k+1)^d} \left((k+1)^d \left(1 - \frac{(k+1)^2\ell^2}{3\tau(M)^2}\right)^d - k^d \left(1 + \frac{4k^2\ell^2}{3\tau(M)^2}\right)^d\right)$$

$$\geq \frac{c_d f_{\min}}{f_{\max}(k+1)^d} \left(dk^{d-1} \left(1 - \frac{(k+1)^2\ell^2}{3\tau(M)^2}\right)^d - \right.$$

$$\left. k^d \left(\left(1 + \frac{4k^2\ell^2}{3\tau(M)^2}\right)^d - \left(1 - \frac{(k+1)^2\ell^2}{3\tau(M)^2}\right)^d\right)\right)$$

$$\geq \frac{c_d f_{\min}}{f_{\max}(k+1)^d} \left(dk^{d-1}/2 - C_4 k^d \frac{k^2\ell^2}{\tau(M)^2}\right) \geq \frac{C_5}{k}$$

where we used the inequality $C_4 \frac{k^2\ell^2}{\tau(M)^2} \leq dk^{-1}/4$ at the last line: this inequality holds as ℓ^2 is of order $(\ln n/n)^{4/d}$ and k^{-3} is at least of order $(\ln n/n)^{3/d}$.

If $Y_1, \ldots, Y_N$ are i.i.d. random variables of law μ, conditioned on being in B_i, we therefore have

$$\mathbb{P}\left(\forall j \in \{1, \ldots, N\}, |Y_j - X_i| \notin [k\ell, (k+1)\ell] | X_i\right) \leq \exp\left(-C_5 N/k\right)$$

For each ball B_i, we let $J_i \subseteq \{L+1, \ldots, n\}$ be the set of indexes $j > L$ such that $X_j \in B_i$. Assume that there exists a set of A balls $B_{i_1}, \ldots, B_{i_A}$ that are pairwise disjoint. Then, the corresponding sets J_i are also pairwise disjoint. Conditionally on $\mathcal{X}_n^0$ and on $N_a := |J_{i_a}|$, the sets $\{X_j : j \in J_{i_a}\}$ are independent for $a = 1, \ldots, A$, and each consists of a sample of N_{i_a} independent points sampled according to μ conditioned on being in B_{i_a}. Therefore, if E is the event that $\mathrm{Rad}(\mathcal{X}_n)$ does not intersect I_k, we have

$$\mathbb{P}(E | \mathcal{X}_n^0, (N_i)_{i \leq L}) \leq \exp\left(-\sum_{a=1}^{A} N_{i_a} C_5/k\right).$$

The random variable $\sum_{a=1}^{A} N_{i_a}$ is the number of points of $\mathcal{X}_n^1$ in $\bigcup_{a=1}^{A} B_{i_a}$. It follows a binomial disribution of parameters $n - L$ and $p = \sum_{a=1}^{A} \mu(B_{i_a}) \geq C_6 A(k\ell)^d$, so that we have

$$\begin{aligned}
\mathbb{P}\left(E | \mathcal{X}_n^0\right) &\leq \mathbb{E}\left[\exp\left(-\sum_{a=1}^{A} N_{i_a} C_5/k\right) \bigg| \mathcal{X}_n^0\right] \\
&\leq \exp\left(-C_6(n-L)A(k\ell)^d \left(1 - e^{-C_5/k}\right)\right) \\
&\leq \exp\left(-C_7(n-L)A(k\ell)^d/k\right).
\end{aligned}$$

The quantity A can be chosen equal to the maximal number of balls B_i that are pairwise disjoint. A procedure to create a set of pairwise disjoint balls is the following. Start with $X_{i_1} = X_1$, and throw away all the points of $\mathcal{X}_n^0$ at distance less than $2(k+1)\ell$ from X_1. Take any point X_{i_2} that has not been thrown away, and throw away all the remaining points that are distance less than $2(k+1)\ell$ from X_{i_2}. Repeating this procedure for $\tilde{A}$ steps until no points are left, we obtain a set of indexes for which the corresponding balls are pairwise disjoint. In particular, $\tilde{A} \leq A$. The number of points that are thrown away at the step a follows a binomial distribution of parameters m and q, where $m \leq L$ is the number of points in $M_a := M \setminus \bigcup_{a' < a} \mathcal{B}\left(X_{i_{a'}}, 2(k+1)\ell\right)$, and, as long as $f_{\max} c_d a(k\ell)^d \leq 1/2$

$$q = \frac{\mu(B_{i_a})}{\mu(M_a)} \leq \frac{c_d f_{\max}(k\ell)^d}{1 - a c_d f_{\max}(k\ell)^d} \leq C_8(k\ell)^d.$$

In particular, the number of points that have been thrown away after a steps is stochastically dominated by the sum of a independent binomial random variables of parameter L and $C_8(k\ell)^d$, that is a binomial random variable Z_a of parameters aL and $C_8(k\ell)^d$. This implies that

$$\mathbb{P}(A \leq a) \leq \mathbb{P}(\tilde{A} \leq a) \leq \mathbb{P}(Z_a \geq L).$$

Let $a = \lfloor 1/\left(C_9(k\ell)^d\right)\rfloor$, where C_9 is choosen so that $f_{\max} c_d a(k\ell)^d \leq 1/2$ and $\mathbb{E}Z_a = aLC_8(k\ell)^d \leq L/2$. Then, $\mathbb{P}(Z_a \geq L) \leq \mathbb{P}(Z_a - \mathbb{E}Z_a \geq L/2) \leq \exp(-C_{10}L)$ using Bernstein's inequality. Therefore, letting $L = (3/C_{10})(\ln n)$, we obtain that

$$\begin{aligned}
\mathbb{P}(E) &\leq \mathbb{E}\left[\exp\left(-C_7(n-L)A(k\ell)^d/k\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(-C_7(n-L)A(k\ell)^d/k\right)1\{A \geq a\}\right] + \mathbb{P}(A \leq a) \\
&\leq \exp\left(-C_{11}n\left\lfloor 1/\left(C_9(k\ell)^d\right)\right\rfloor (k\ell)^d/k\right) + n^{-3} \\
&\leq \exp\left(-C_{12}n^{1-1/d}(\ln n)^{1/d}/(2\beta)\right) + n^{-3}
\end{aligned}$$

If $d \geq 2$, the first term in the above sum is smaller than n^{-3}. If $d = 1$, it is equal to $n^{-C_{12}/(2\beta)}$ and we choose $\beta = C_{12}/6$ to conclude. □

Proof of the lower bound in (26)

A first naive attempt to lower bound $t_\lambda(\mathcal{X}_n)$ is the following. Remark that if two points X_1, X_2, at distance $2t$, are such that the ball centered at their middle, of radius t, does not contain any point of $\mathcal{X}_n$, then $d_H(\mathrm{Conv}(\{X_1, X_2\})|\mathcal{X}_n) = t$. Fix $t > 0$, and assume that there is some $t' \in \mathrm{Rad}(\mathcal{X}_n)$ smaller than t such that $h(t', \mathcal{X}_n) < t'$. There must then exist a simplex of size at least 3 of radius smaller than t in $\mathcal{X}_n$. In particular, there are three points X_1, X_2 and X_3 of $\mathcal{X}_n$ so that $X_2, X_3 \in \mathcal{B}(X_1, 2t)$. Therefore, according to Lemma 2.3, if $t \leq \tau(M)/8$,

$$\begin{aligned}
\mathbb{P}\left(t_\lambda(\mathcal{X}_n) < t\right) &\leq \mathbb{P}\left(\exists X_1, X_2, X_3 \text{ with } X_2, X_3 \in \mathcal{B}(X_1, 2t)\right) \\
&\leq \mathbb{E}\left[\mathbb{P}\left(\exists X_2, X_3 \in \mathcal{B}(X_1, 2t)|X_1\right)\right] \\
&\leq \mathbb{E}\left[(n\mu(\mathcal{B}(X_1, 2t))^2\right] \leq (\alpha_d f_{\max} n(13t/6)^d)^2 \leq C_0(nt^d)^2.
\end{aligned}$$
(28)

We know from the previous section that $t^*(\mathcal{X}_n)$ is of order $t \simeq (\ln n/n)^{1/d}$, while $\left(nt^d\right)^2 \simeq (\ln n)^2$ for such a value of t. Hence, the previous inequality is far from sufficient to obtain Theorem 4.6. We therefore consider a more elaborate construction.

Lemma 4.9. *Let $\delta > 0$. For t small enough (depending on μ and δ), there exist K pairwise disjoint measurable subsets $U_1, \ldots, U_K$, so that $K \geq c_{u,\delta}t^{-d}$ and each set U_k contains a ball V_k of radius t and satisfies*

$$\mu(U_k) = m(t) := \alpha_d(1+\delta)f_{\min}t^d. \tag{29}$$

Before proving the lemma, note that we also have $Km(t) \leq 1$ by a union bound.

Proof. Consider the collection $\mathcal{F}$ of balls V of radius t centered at a point of M satisfying $\mu(V) \leq \alpha_d(1+\delta)f_{\min}t^d$, and let A_t be the set of the centers of such balls. By Besicovitch's covering theorem [20, Theorem 2.8.14], there exist N_M collections $\mathcal{G}_1, \ldots, \mathcal{G}_{N_M}$ of disjoint balls in $\mathcal{F}$ such that

$$A_t \subseteq \bigcup_{l=1}^{N_M} \bigcup_{V \in \mathcal{G}_l} V.$$

Letting K_t be the maximal number of pairwise disjoint balls in $\mathcal{F}$, we have $\mu(A_t) \leq N_M K_t \alpha_d (1+\delta) f_{\min} t^d$. By the Lebesgue differentiation theorem, for almost all points $x \in M$ with $f(x) < (1+\delta) f_{\min}$, we have

$$\lim_{t \to 0} \frac{\mu(\mathcal{B}(x,t))}{\alpha_d t^d} < f_{\min}(1+\delta).$$

For such a x, we then have $x \in \liminf_{t \to 0} A_t$. Therefore,

$$c_\mu = \mu(\{x \in M : f(x) < (1+\delta) f_{\min}\}) \leq \mu\left(\liminf_{t \to 0} A_t\right) \leq \liminf_{t \to 0} \mu(A_t)$$

$$\leq N_M \alpha_d (1+\delta) \liminf_{t \to 0} K_t t^d.$$

By the definition of $f_{\min}, c_\mu > 0$. Therefore, for t small enough, we have $K_t \geq \frac{c_\mu}{2 N_M \alpha_d (1+\delta)} t^{-d}$. Let $V_1, \ldots, V_{K(t)}$ be a set of pairwise disjoint balls in $\mathcal{F}$. By construction, each ball V_k satisfies $\mu(V_k) \leq m(t)$. Also, we have $\mu(V_k) \geq \alpha_d f_{\min} t^d / 2$ for t small enough by Lemma 2.3. This implies by a union bound that $1 \geq K(t) \alpha_d f_{\min} t^d / 2$. Therefore, $K(t) m(t) \leq 2(1+\delta)$. We define $K = \lfloor K(t)/(2(1+\delta)) \rfloor$, a number that satisfies $K \geq c_{\mu,\delta} t^{-d}$ and $K m(t) \leq 1$.

Eventually, we build the sets U_k by induction by choosing any measurable set W_k in $M \backslash (\bigcup_{k' < k} U_{k'} \cup V_k)$ with $\mu(W_k) = m(t) - \mu(V_k) \geq 0$ and letting $U_k = V_k \cup W_k$. This is possible as

$$\mu\left(M \backslash (\bigcup_{k' < k} U_{k'} \cup V_k)\right) \geq 1 - (k-1) m(t) - \mu(V_k) \geq m(t) - \mu(V_k).$$

By construction, $\mu(U_k) = m(t)$ for every k. Note that we used the fact that for any $A \subseteq M$ and $0 \leq p \leq \mu(A)$, there exists a subset $V \subseteq A$ with $\mu(V) = p$: this holds as μ is absolutely continuous with respect to the volume measure on M. □

We fix such a partition in the following, with balls V_k of radius $(2-\lambda)t$. We write m for $m((2-\lambda)t)$. Let B_k be the ball sharing its center with V_k, of radius t. For $W \subseteq M$, let $N(W)$ be the number of points of $\mathcal{X}_n$ in W. We also write N_k for $N(U_k)$. Let x_k be the center of B_k and e be a unit vector in $T_x M$, and denote by A_k^+ (resp. A_k^-) the ball of radius $(1-\lambda)t/2$ centered at $x^+ = x_k + e(1+\lambda)t/2$ (resp. $x^- = x_k - e(1+\lambda)t/2$), see Figure 6.

Lemma 4.10. *Fix $k = 1, \ldots, K$. If $h(t, \mathcal{X}_n) < \lambda t$ and $N_k = 2$, then we cannot have both $N(A_k^+) = 1$ and $N(A_k^-) = 1$*

Proof. Let $\sigma = \mathcal{X}_n \cap U_k$. Assume by contradiction that $h(t, \mathcal{X}_n) < \lambda t$, $N_k = 2$, and $N(A_k^+) = N(A_k^-) = 1$. Then, σ is made of two points, x_1 and x_2, respectively in A_k^+ and A_k^-. As both points belong to B_k, we have $r(\sigma) \leq t$. Therefore, $d_H(\text{Conv}(\sigma)|\mathcal{X}_n) \leq h(t, \mathcal{X}_n) < \lambda t$. In particular, the middle point x_0 of x_1 and x_2 is at distance less than λt from $\mathcal{X}_n$. Let us show that $\mathcal{B}_M(x_0, |x_1 - x_0|) \subseteq V_k$. If this is the case, then $d(x_0, \mathcal{X}_n) = |x_1 - x_2|/2 \geq \lambda t$, a contradiction with
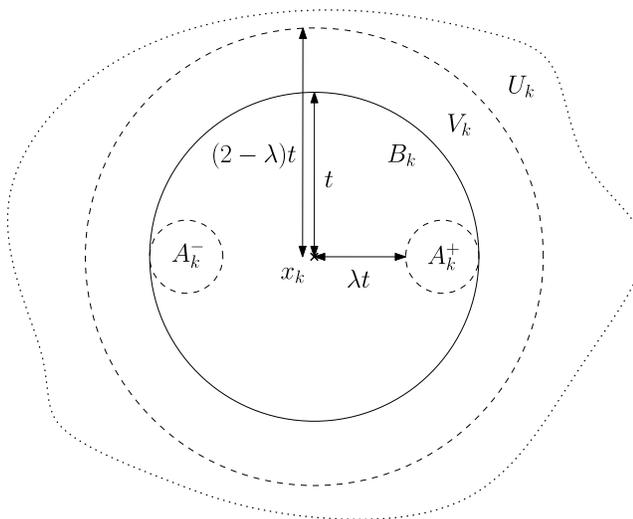
FIG 6. *Any ball with diameter whose one extremity is in A_k^- and the other in A_k^+ is included in U_k.*

having $d_H(\mathrm{Conv}(\sigma)|\mathcal{X}_n) < \lambda t$. Let $z \in \mathcal{B}_M(x_0, |x_1 - x_0|)$ and denote by π_e the projection on e. Then,

$$|z - x_k| \le |z - x_0| + |x_0 - x_k| \le \frac{|x_1 - x_2|}{2} + |\pi_e(x_0 - x_k)| + \left|\pi_e^{\perp}(x_0 - x_k)\right|$$
$$\le t + \frac{(1-\lambda)t}{2} + \frac{(1-\lambda)t}{2} \le (2-\lambda)t$$

concluding the proof. $\qquad\square$

Denote by F_k the complementary event of the event $N(A_k^+) = N(A_k^-) = 1$. We obtain the bound

$$\mathbb{P}(h(t, \mathcal{X}_n) < \lambda t) \le \mathbb{P}(\forall k = 1, \ldots, K, N_k \ne 2 \text{ or } (N_k = 2 \text{ and } F_k))$$
$$= \mathbb{E}\left[\mathbb{P}\left(\forall k = 1, \ldots, K, N_k \ne 2 \text{ or } (N_k = 2 \text{ and } F_k) \,\middle|\, (N_k)_{k=1,\ldots,K}\right)\right]$$
$$\le \mathbb{E}\left[\prod_{k=1}^{K}(\mathbf{1}\{N_k \ne 2\} + \mathbb{P}(F_k|N_k = 2)\,\mathbf{1}\{N_k = 2\})\right]$$
$$\le \mathbb{E}\left[\prod_{k=1}^{K}(1 - (1 - \mathbb{P}(F_k|N_k = 2))\,\mathbf{1}\{N_k = 2\})\right].$$

Lemma 4.11. *There exists a positive constant C_1 such that*

$$\mathbb{P}(F_k|N_k = 2) \le e^{-C_1} \text{ for } k = 1, \ldots, K. \tag{30}$$

Proof. If $|x_+ - x_k| \le t \le 7\tau(M)/24$, then there exists $y_+ \in M$ that satifies $\pi_{x_k}(y_+ - x_k) = x_+ - x_k$ by Lemma 2.2. Furthermore, we have $|y_+ - x_k| \le 8t/7$ and, by Lemma 2.1, we have $|y_+ - x_+| \le (8t/7)^2/(2\tau(M)) = 32t^2/(49\tau(M))$. In particular,

$$\mathcal{B}\left(x_+, (1-\lambda)t/2\right) \supseteq \mathcal{B}\left(y_+, (1-\lambda)t/2 - 32t^2/(49\tau(M))\right) \supseteq \mathcal{B}\left(y_+, (1-\lambda)t/4\right)$$

if $t \le 49(1-\lambda)\tau(M)/128$. According to Lemma 2.3, we therefore have, also assuming that $t \le \tau(M)/4$

$$\mu\left(\mathcal{B}\left(x_+, (1-\lambda)t/2\right)\right) \ge f_{\min}\alpha_d\left(\frac{(1-\lambda)t}{4}\frac{47}{48}\right)^d$$

and the same inequality holds for x_-.

Let Y_1, Y_2 be two independent random variables sampled according to μ, conditioned on being in U_k. Then, as $\mu(U_k) = m = \alpha_d(1+\delta)f_{\min}(2-\lambda)^d t^d$,

$$\begin{aligned}
\mathbb{P}\left(F_k | N_k = 2\right) &= 1 - 2\mathbb{P}\left(Y_1 \in A_k^+\right)\mathbb{P}\left(Y_2 \in A_k^-\right) \\
&= 1 - 2\frac{\mu\left(\mathcal{B}\left(x^+, (1-\lambda)t/2\right)\right)\mu\left(\mathcal{B}\left(x^-, (1-\lambda)t/2\right)\right)}{\mu(U_k)^2} \\
&\le 1 - 2\left(\frac{\left(\frac{47}{48}\frac{1-\lambda}{4}\right)^d}{(1+\delta)(2-\lambda)^d}\right)^2 \le e^{-C_1}
\end{aligned}$$

where $C_1 = 2\left(\frac{\left(\frac{47}{48}\frac{1-\lambda}{4}\right)^d}{(1+\delta)(2-\lambda)^d}\right)^2$. $\qquad\square$

We finally obtain

$$\mathbb{P}\left(h(t, \mathcal{X}_n) < \lambda t\right) \le \mathbb{E}\left[\exp\left(-C_1\sum_{k=1}^{K}\mathbf{1}\left\{N_k = 2\right\}\right)\right]. \tag{31}$$

Lemma 4.12. *Assume that $nm \le \max\left(m^{-1}, (\ln n)^2\right)$. Let $\phi : x \in [0, +\infty) \mapsto \min(1, x)e^{-x}$. Then,*

$$\mathbb{E}\left[\exp\left(-C_1\sum_{k=1}^{K}\mathbf{1}\left\{N_k = 2\right\}\right)\right] \le C_2\exp\left(-C_3 n\phi(nm)\right) \tag{32}$$

for some positive constants C_2, C_3.

Lemma 4.12 relies on concentration inequalities and is proved in Section A. As m is of order t^d, the condition $nm \le \max\left(m^{-1}, (\ln n)^2\right)$ is satisfied as long as $t^d \ll (\ln n)^2/n$. Remark also that the function ϕ is increasing on $[0, 1]$ and decreasing on $[1, +\infty)$.

Assume that $t_1 \le t \le t_2$, where

$$t_1 = \frac{1}{2-\lambda} \left(\frac{1}{\alpha_d f_{\min}(1+\delta)n} \right)^{1/d} \text{ and } t_2 = \frac{1}{2-\lambda} \left(\frac{\beta \ln n}{\alpha_d f_{\min}(1+\delta)n} \right)^{1/d}$$

for some $0 < \beta < 1$. Then, $1 \leq nm \leq \beta \ln n$, so that $\phi(nm) \geq \phi(\beta \ln n) = n^{-\beta}$ and

$$\forall t \in [t_1, t_2], \quad \mathbb{P}\left(h(t, \mathcal{X}_n) < \lambda t\right) \leq C_2 \exp\left(-C_3 n^{1-\beta}\right) \leq n^{-2} \tag{33}$$

for n large enough. Assume now that $t \in [t_0, t_1]$, where

$$t_0 = \frac{1}{2-\lambda} \left(\frac{\kappa \ln n}{\alpha_d f_{\min}(1+\delta)n^2} \right)^{1/d}$$

for some $\kappa > 0$. Then, $\kappa(\ln n/n) \leq nm \leq 1$, so that $\phi(nm) \geq \phi(\kappa(\ln n/n)) \geq \kappa \ln n/(2n)$ for n large enough. Choosing $\kappa \geq 4/C_3$, we obtain that

$$\forall t \in [t_0, t_1], \quad \mathbb{P}\left(h(t, \mathcal{X}_n) < \lambda t\right) \leq C_2 n^{-C_3 \kappa/2} \leq C_2 n^{-2}. \tag{34}$$

The picture is now as follows. We know from (28) that $t_\lambda(\mathcal{X}_n) \geq t_0$ with probability at least $1 - c_3(\ln n/n)^2$. For each t between t_0 and t_2, we also have $h(t, \mathcal{X}_n) \geq \lambda t$ with probability at least n^{-2} (at least for n large enough with respect to λ and μ). Consider a sequence $t^{(i)}$ with $t^{(0)} = t_0$ and $t^{(i+1)} = t^{(i)}/\lambda$ for $i = 0, \ldots, I$, with I chosen so that

$$t_2/\lambda \leq t^{(I)} \leq t_2$$

Assume that $t_\lambda(\mathcal{X}_n) \geq t_0$ and that $h(t^{(i)}, \mathcal{X}_n) \geq \lambda t^{(i)}$ for every i. If t belongs to the interval $\left[t^{(i)}, t^{(i+1)}\right]$, then $h(t, \mathcal{X}_n) \geq h(t^{(i)}, \mathcal{X}_n) \geq \lambda t^{(i)} \geq \lambda^2 t$. Therefore, $t_{\lambda^2}(\mathcal{X}_n) \geq t_2$ Let $\lambda' = \lambda^2$. As I is of order $\ln n$, by a union bound, we obtain that, for any $0 < \beta, \delta < 1$, $\lambda' \in (0,1)$ and n large enough

$$\mathbb{P}\left(t_{\lambda'}(\mathcal{X}_n) \leq \frac{1}{2-\sqrt{\lambda'}} \left(\frac{\beta \ln n}{\alpha_d f_{\min}(1+\delta)n} \right)^{1/d} \right)$$

$$\leq \mathbb{P}\left(t_\lambda(\mathcal{X}_n) < t_0\right) + \sum_{i=0}^{I} \mathbb{P}\left(h(t^{(i)}, \mathcal{X}_n) < \lambda t^{(i)} \right) \tag{35}$$

$$\leq c_3(\ln n/n)^2 + c_4(\ln n)n^{-2}$$

$$\leq c_5(\ln n/n)^2.$$

Lemma 4.13. *Let $A \subseteq M$. Let $0 < \lambda \leq \lambda' < 1$. Then, $t_\lambda(A) \geq \frac{\lambda'}{\lambda} t_{\lambda'}(A)$.*

Proof. The function $h(\cdot, A)$ is nondecreasing, and is therefore larger than $\lambda' t_{\lambda'}(A)$ for $t \geq t_{\lambda'}(A)$. Therefore, for $t \in [t_{\lambda'}(A), (\lambda'/\lambda) t_{\lambda'}(A)]$, we have $h(t, A) \geq \lambda' t_{\lambda'}(A) \geq \lambda t$, yielding the conclusion. $\square$

Let $0 < \lambda < (1+b)^{-1/d}$. From Proposition 3.9, we know that with probability $1 - c(\ln n)^{d-1} n^{-b}$, we have $\varepsilon(\mathcal{X}_n) \leq (1+b)^{1/d} (\ln n / (n\alpha_d f_{\min}))^{1/d}$. For any $r > 1$, if n is large enough, by Proposition 3.4, this entails that $t^*(\mathcal{X}_n) \leq r\varepsilon(\mathcal{X}_n)$. Choose λ', β and r close enough to 1, and δ small enough, so that

$$\frac{\lambda'}{\lambda \left(2 - \sqrt{\lambda'} \right)} \frac{\beta^{1/d}}{(1+\delta)^{1/d}} \geq r(1+b)^{1/d}.$$

Such a choice is possible as $\frac{1}{\lambda} > (1+b)^{1/d}$. Then, assuming that the complementary of the event described in (35) also holds, we have

$$t_\lambda(\mathcal{X}_n) \geq \frac{\lambda'}{\lambda} t_{\lambda'}(\mathcal{X}_n) \geq \frac{\lambda'}{\lambda} \frac{1}{2 - \sqrt{\lambda'}} \left(\frac{\beta \ln n}{\alpha_d f_{\min}(1+\delta)n} \right)^{1/d} \geq r\varepsilon(\mathcal{X}_n) \geq t^*(\mathcal{X}_n)$$

As the probability appearing in (35) is smaller than a quantity of order $(\ln n)^2 n^{-2} \leq (\ln n)^a n^{-b}$ for any $0 < b \leq 2$, we obtain inequality (26), concluding the proof of the first statement of Theorem 4.6.

Proof of (26)

Consider a n-sample $\{X_1, \ldots, X_n\}$ on the circle M of radius 1. Without loss of generality, we assume that $X_1 = (0, 1)$. Each point X_i is equal to $\exp(i\theta_i)$ where $\theta_i \in [0, 2\pi)$. Consider the ordering

$$0 = \theta_{(1)} \leq \cdots \leq \theta_{(n)}$$

and the associated points $X_{(1)}, \ldots, X_{(n)}$. Define the spacings $V_i = \theta_{(i+1)} - \theta_{(i)}$ for $i = 1, \ldots, n$ (with by convention $\theta_{(n+1)} = 2\pi$). The corresponding edge lenth $\left| X_{(i+1)} - X_{(i)} \right| =: 2t_i$ satisfies $V_i = \arccos\left(1 - 2t_i^2\right)$.

We write $V_{(1)} \leq \cdots \leq V_{(n)}$ for the ordered spacings (and $t_{(1)} \leq \cdots \leq t_{(n)}$ for the associated lengths). Note that we have $t^*(\mathcal{X}_n) = t_{(n)}$. The next lemma asserts that the convexity defect function cannot increase too much between two consecutive $t_{(i)}$s.

Lemma 4.14. *For $t \in \left[t_{(i)}, t_{(i+1)} \right)$, we have $h(t, \mathcal{X}_n) \leq t_{(i)} + t_{(i+1)}^2$.*

Proof. Let $\left[X_{(k)}, X_{(l)} \right]$ be an edge of length smaller than $2t$ with $k < l$. We assume without loss of generality that X_0 does not lie on the arc between $X_{(k)}$ and $X_{(l)}$. Let x be a point on this edge, of the form $re^{i\theta}$ for some angle $\theta_{(k)} \leq \theta \leq \theta_{(l)}$. The angle θ belongs to the segment $\left[\theta_{(j)}, \theta_{(j+1)} \right]$ for some index j. As $t < t_{(i+1)}$ we have $t_j < t_{(i+1)}$, that is $t_j \leq t_{(i)}$. The ray of angle θ hits the line $\left[X_{(j)}, X_{(j+1)} \right]$ at some point y, and

$$d(x, \mathcal{X}_n) \leq |x - y| + d(y, \mathcal{X}_n) \leq d(x, M) + t_j \leq t^2 + t_{(i)}$$

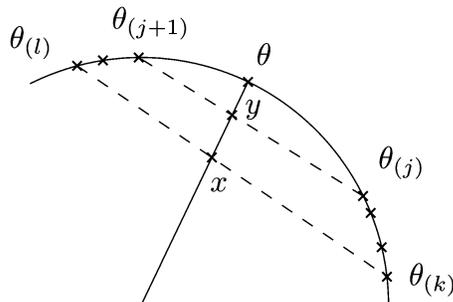by Lemma 3.2. As $t \leq t_{(i+1)}$, we obtain the conclusion. $\qquad\square$

FIG 7. *Construction used in the proof of Lemma 4.14. The distance between $X_{(j)} = e^{i\theta_{(j)}}$ and $X_{(j+1)} = e^{i\theta_{(j+1)}}$ is equal to $2t_j$, while the distance between $X_{(k)} = e^{i\theta_{(k)}}$ and $X_{(l)} = e^{i\theta_{(l)}}$ is smaller than $2t$.*

Let $\lambda > (1 + b)^{-1}$ and fix an arbitrary λ' satisfying $(1 + b)^{-1} < \lambda' < \lambda$. Assume that $t_{(n-1)} \leq \lambda' t_{(n)}$. Then, for $t \in [t_{(n-1)}, t_{(n)})$, we have according to the previous lemma that

$$h(t, \mathcal{X}_n) \leq \lambda' t_{(n)} + t_{(n)}^2.$$

Choosing $t \in I := [t_{(n)} \left(\lambda' + t_{(n)}\right) / \lambda, t_{(n)})$, we have $h(t, \mathcal{X}_n) < \lambda t$. The interval I satisfies the conditions of Lemma 4.7, and therefore intersects $\mathrm{Rad}(\mathcal{X}_n)$ with probability $1 - n^{-2}$. In particular, $t_\lambda(\mathcal{X}_n)$ is smaller than the upper endpoint of I, that is $t_{(n)} = t^*(\mathcal{X}_n)$. Note that such a choice of t is possible as long as $\lambda - \lambda' > t_{(n)}$. To put it another way, we have

$$\mathbb{P}\left(t_\lambda(\mathcal{X}_n) < t^*(\mathcal{X}_n)\right) \leq \mathbb{P}\left(t_{(n-1)} \leq \lambda' t_{(n)}\right) + \mathbb{P}\left(\lambda - \lambda' < t_{(n)}\right) + n^{-2}. \qquad (36)$$

The second probability in the above equation is exponentially small by Proposition 3.9. It remains to study the probability that $t_{(n-1)} \leq \lambda' t_{(n)}$. Let A_1, $\ldots, A_n$ be a n-sample following an exponential distribution. According to [18, Section 6.4], we have

$$(V_1, \ldots, V_n) \sim 2\pi \left(\frac{A_1}{\sum_{i=1}^n A_i}, \ldots, \frac{A_n}{\sum_{i=1}^n A_i} \right)$$

In particular, the law of $V_{(n)}/V_{(n-1)}$ is equal to the law of $A_{(n)}/A_{(n-1)}$, the largest of the A_is divided by the second largest. Furthermore, according to [37, Theorem 2.1], we have, for any $s > 1$,

$$\mathbb{P}\left(A_{(n)}/A_{(n-1)} \ge s\right) = n(n-1)\sum_{k=0}^{n-2}\binom{n-2}{k}\frac{(-1)^{n-2-k}}{n-1-k+s}$$

$$= n(n-1)\sum_{k=0}^{n-2}\binom{n-2}{k}(-1)^{n-2-k}\int_0^1 x^{n-2-k+s}\,\mathrm{d}x$$

$$= n(n-1)\int_0^1 x^s(1-x)^{n-2}\,\mathrm{d}x$$

$$= n(n-1)B(s+1,n-1)$$

$$\sim n^2\Gamma(s+1)n^{-(s+1)} \sim \Gamma(s+1)n^{1-s},$$

$$\tag{37}$$

where B is the Beta function. Also, by writing a Taylor expansion of arccos at 1, we obtain that for $t_{(n)}$ small enough,

$$\frac{V_{(n)}}{V_{(n-1)}} = \frac{\arccos\left(1 - 2t_{(n)}^2\right)}{\arccos\left(1 - 2t_{(n-1)}^2\right)} \le \frac{t_{(n)}}{t_{(n-1)}}\left(1 + \frac{5t_{(n)}^2}{24}\right)$$

If $t_{(n-1)} \le \lambda' t_{(n)}$, then we have $\frac{V_{(n)}}{V_{(n-1)}} \ge (\lambda')^{-1}\left(1 + \frac{5t_{(n)}^2}{24}\right) \ge 1 + b$ if $t_{(n)}$ is smaller than some constant c_0 (recall that $\lambda' > (1+b)^{-1}$). Therefore,

$$\mathbb{P}\left(t_{(n-1)} \le \lambda' t_{(n)}\right) \ge \mathbb{P}\left(t_{(n)} > c_0\right) + \mathbb{P}\left(\frac{V_{(n)}}{V_{(n-1)}} \ge 1 + b\right)$$

for some small constant c_0 (depending on the distance between $(1+b)^{-1}$ and λ'). The first probability is exponentially small, and the second one is of order n^{-b} by (37). Inequality (36) then yields the conclusion.

5. Sampling with noise

So far, we have always considered that the point cloud $\mathcal{X}_n$ lies exactly on the manifold M. However, all the constructions presented are stable with respect to tubular noise.

Let $0 < \gamma < \tau_{\min}$. Let $X = Y + Z$, with the law ν of Y being in $\mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}}$ and $Z \in T_Y M^\perp$ satisfying $|Z| \le \gamma$. We let $\mathcal{Q}^{d,\gamma}_{\tau_{\min}, f_{\min}, f_{\max}}$ be the set of laws of such random variables X. Observe that, as we do not assume that the conditional noise $Z|Y$ is centered, the model is not identifiable, that is M is not determined by the law μ of X. To simplify matters, for each law $\mu \in \mathcal{Q}^{d,\gamma}_{\tau_{\min}, f_{\min}, f_{\max}}$, we will make an arbitrary choice among the admissible couples (Y, Z) with $Y + Z \sim \mu$. The "underlying manifold M of the law μ" will be the support of the law of Y, while the results of this section will hold for any choice of couple (Y, Z).

Remark 5.1 (On the orthogonality assumption). *The assumption that the noise is orthogonal (that is $Z \in T_Y M^\perp$) is not restrictive. Let $\gamma < \tau_{\min}$, $\nu \in \mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}}$ with density f and $Y \sim \nu$. Let Z be any random variable supported on $\mathcal{B}(0, \gamma)$, and $X = Y + Z$ (without necessarily $Z \in T_Y M^\perp$). We may write $X = \pi_M(X) + (X - \pi_M(X)) = Y' + Z'$. By Lemma 2.2, we have $Z' \in T_Y, M^\perp$. Furthermore, the density of Y' can be explicitly computed in terms of the density of f and of the Jacobian of the function $G_z : y \in M \mapsto \pi_M(y + z)$. More precisely, one can show that G_z is bijective, of class $\mathcal{C}^1$, and, by a change of variable, that the density f' of Y' at y is given by $\mathbb{E}\left[f(G_Z^{-1}(y))J(G_Z^{-1})(y)\right]$. The derivative of G_Z is expressed in terms of the second fundamental form of M (whose operator norm is bounded by the reach $\tau(M)$ [31]). In particular, the Jacobian is upper and lower bounded, so that f' is lower and upper bounded on M. In other words, the law of X belongs to $\mathcal{Q}^{d,\gamma}_{\tau_{\min}, a f_{\min}, f_{\max}/a}$ for some $0 < a < 1$ depending on $d, \tau_{\min}$ and $\tau_{\min} - \gamma$.*

We first show that the t-convex hull with parameter t of order $(\ln n/n)^{1/d}$ has a risk of the same order if tubular noise is added.

Proposition 5.2. *Let $A, B \subseteq \mathbb{R}^D$ and let $d_H(A, B) \leq \gamma$. Then,*

$$d_H(\mathrm{Conv}(t, A)| \mathrm{Conv}(t + \gamma, B)) \leq \gamma \tag{38}$$

Proof. Let $\sigma \subseteq A$. By definition, there exists $\sigma' \subseteq B$ such that $d_H(\sigma|\sigma') \leq \gamma$. We have $r(\sigma') \leq r(\sigma) + \gamma \leq t + \gamma$ (see [10, Lemma 16]) and $d_H(\mathrm{Conv}(\sigma)| \mathrm{Conv}(\sigma')) \leq \gamma$. ∎

Let $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ be a n-sample of law μ, with $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$ the corresponding sample on M (that is $Y_i = \pi_M(X_i)$. If $t \geq t^*(\mathcal{Y}_n) + \gamma$, then

$$\begin{aligned}
d_H(M| &\mathrm{Conv}(t, \mathcal{X}_n)) \\
&\leq d_H(M| \mathrm{Conv}(t - \gamma, \mathcal{Y}_n)) + d_H(\mathrm{Conv}(t - \gamma, \mathcal{Y}_n)| \mathrm{Conv}(t, \mathcal{X}_n)) \\
&\leq \frac{(t - \gamma)^2}{\tau(M)} + \gamma
\end{aligned}$$

and $\quad d_H(\mathrm{Conv}(t, \mathcal{X}_n)|M)$

$$\begin{aligned}
&\leq d_H(\mathrm{Conv}(t, \mathcal{X}_n)| \mathrm{Conv}(t + \gamma, \mathcal{Y}_n)) + d_H(\mathrm{Conv}(t + \gamma, \mathcal{Y}_n)|M) \\
&\leq \gamma + \frac{(t + \gamma)^2}{\tau(M)}.
\end{aligned}$$

Therefore, we obtain that, for $t \geq t^*(\mathcal{Y}_n) + \gamma$,

$$d_H(\mathrm{Conv}(t, \mathcal{X}_n), M) \leq \frac{(t + \gamma)^2}{\tau(M)} + \gamma. \tag{39}$$

Assume that $\gamma \leq \eta(\ln n/n)^{2/d}$ for some $\eta > 0$ and let $\tilde{t}_n = 2t_n$, where t_n is the radius appearing in Theorem 1.1. The probability that $\tilde{t}_n \leq t^*(\mathcal{Y}_n) + \gamma$ is smaller than the probability that $t_n \leq t^*(\mathcal{Y}_n)$, a probability that we control by Proposition 3.9. As $\tilde{t}_n + \gamma \leq 3t_n$ for n large enough, we obtain that

$$\mathbb{E}\left[d_H(\mathrm{Conv}(\tilde{t}_n, \mathcal{X}_n), M)\right] \leq \left(\frac{c_1}{\tau_{\min}\left(\alpha_d f_{\min}\right)^{2/d}} + \eta\right)\left(\frac{\ln n}{n}\right)^{2/d}$$

for some absolute constant c_1.

Let us now analyze how the selection procedure is impacted by the presence of noise. We mimick the proof of Theorem 4.6. Let $0 < b \leq 2$ and let $0 < \lambda < (1+b)^{-1/d}$. If $t_\lambda(\mathcal{X}_n) < t$, then in particular there exist three points X_1, X_2 and X_3 such that $X_2, X_3 \in \mathcal{B}(X_1, 2t)$. We then have by Lemma 2.2 that $Y_2, Y_3 \in \mathcal{B}\left(Y_1, \frac{2\tau(M)}{\tau(M)-\gamma}t\right)$. We obtain as in (28) that

$$\mathbb{P}\left(t_\lambda(\mathcal{X}_n) < t\right) \leq C_1 \frac{\left(nt^d\right)^2}{\left(\tau_{\min} - \gamma\right)^2}. \tag{40}$$

Fix $t \in [t_0, t_2]$ (where t_0 and t_2 are defined in the proof of Theorem 4.6) and let $0 < \gamma < t$. We have, by Proposition 5.2,

$$
\begin{aligned}
h(t, \mathcal{X}_n) &= d_H(\mathrm{Conv}(t, \mathcal{X}_n)|\mathcal{X}_n) \\
&\geq d_H(\mathrm{Conv}(t-\gamma, \mathcal{Y}_n)|\mathcal{Y}_n) - d_H(\mathcal{X}_n|\mathcal{Y}_n) - d_H(\mathrm{Conv}(t-\gamma, \mathcal{Y}_n)|\mathrm{Conv}(t, \mathcal{X}_n)) \\
&\geq h(t-\gamma, \mathcal{Y}_n) - 2\gamma.
\end{aligned}
$$

Therefore, if $\frac{\lambda t + 2\gamma}{t-\gamma} \leq \lambda' < 1$ and $h(t-\gamma, \mathcal{Y}_n) \geq \lambda'(t-\gamma)$, then $h(t, \mathcal{X}_n) \geq \lambda t$. Assume that $\gamma \leq \eta(\ln n/n)^{2/d}$ for some $\eta > 0$ and fix $\lambda' = (1+\lambda)/2$. Then, for $t \geq \tilde{t}_0 := 6\gamma/(1-\lambda)$, the condition $\frac{\lambda t + \gamma}{t-\gamma} \leq \lambda'$ is satisfied. Furthermore, according to the proof of Theorem 4.6, for such a t, the condition $h(t-\gamma, \mathcal{Y}_n) \geq \lambda'(t-\gamma)$ is satisfied with probability at least $1 - cn^{-2}$. Using the same argument than in the proof of Theorem 4.6, we then obtain that

$$
\begin{aligned}
\mathbb{P}\left(t_\lambda(\mathcal{X}_n) \leq \frac{1}{2-\sqrt{\lambda}}\left(\frac{\beta \ln n}{\alpha_d f_{\min}(1+\delta)n}\right)^{1/d}\right) &\leq c_1(\ln n)n^{-2} + c_2\left(n\tilde{t}_0^d\right)^2 \\
&\leq 2c_2 \frac{(\ln n)^4}{n^2}.
\end{aligned}
\tag{41}
$$

We may conclude as in the previous proof that we have $t_\lambda(\mathcal{X}_n) \geq t^*(\mathcal{Y}_n) + \gamma$ with probability equal to $1 - c(\ln n)^{\tilde{a}}n^{-b}$, where $\tilde{a} = 4 \vee (d-1)$ if $b = 2$ and $d-1$ otherwise.

Let us now provide an upper bound on $t_\lambda(\mathcal{X}_n)$. Consider the interval $I = [(1-\lambda/8)t_{\lambda/2}(\mathcal{Y}_n), (1-\lambda/16)t_{\lambda/2}(\mathcal{Y}_n))$. By Theorem 4.6, Proposition 3.4, Lemma 4.8 and Proposition 3.9, $t_{\lambda/2}(\mathcal{Y}_n)$ is at least of order $n^{-1/d}$ and at most of order $(\ln n/n)^{1/d}$ with probability $1 - n^{-2}$. By Lemma 4.7, this implies that $\mathrm{Rad}(\mathcal{Y}_n)$ intersects I with the same probability. Let $t' \in \mathrm{Rad}(\mathcal{Y}_n) \cap I$. This scale corresponds to some simplex $\sigma' = \{y_1, \ldots, y_K\}$, and we let $\sigma = \{x_1, \ldots, x_K\} \subseteq \mathcal{X}_n$ where $y_i = \pi_M(x_i)$. We have $t := r(\sigma) \leq \gamma + t'$ according to [10, Lemma 16]. Furthermore, if z is the center of the smallest enclosing ball of σ, we have

using Lemma 2.2, $|y_i - \pi_M(z)| \leq \frac{\tau(M)}{\tau(M)-\gamma}|x_i - z| \leq \frac{t\tau(M)}{\tau(M)-\gamma}$, indicating that $t' \leq \frac{t\tau(M)}{\tau(M)-\gamma}$. Recalling that γ is of order $(\ln n/n)^{2/d} \ll n^{-1/d} \ll t_{\lambda/2}(\mathcal{Y}_n)$, this means we have found a scale $t \in \mathrm{Rad}(\mathcal{X}_n)$ satisfying

$$(1 - \lambda/4)t_{\lambda/2}(\mathcal{Y}_n) \leq \left(1 - \frac{\gamma}{\tau(M)}\right)t' \leq t \leq t' + \gamma \leq (1 - \lambda/8)t_{\lambda/2}(\mathcal{Y}_n). \quad (42)$$

Using Proposition 4.3 and (42), we obtain

$$
\begin{aligned}
h(t&, \mathcal{X}_n)\\
&\leq d_H(\mathrm{Conv}(t, \mathcal{X}_n)|\mathrm{Conv}(t_{\lambda/2}(\mathcal{Y}_n), \mathcal{Y}_n)) + h(t_{\lambda/2}(\mathcal{Y}_n), \mathcal{Y}_n) + d_H(\mathcal{X}_n, \mathcal{Y}_n)\\
&\leq (t_{\lambda/2}(\mathcal{Y}_n) - t) + \frac{\lambda}{2}t_{\lambda/2}(\mathcal{Y}_n) + \gamma\\
&\leq \frac{\lambda}{4}t_{\lambda/2}(\mathcal{Y}_n) + \frac{\lambda}{2}t_{\lambda/2}(\mathcal{Y}_n) + \gamma \leq \frac{3\lambda}{4}t_{\lambda/2}(\mathcal{Y}_n) + \gamma \leq \lambda(1 - \lambda/4)t_{\lambda/2}(\mathcal{Y}_n)\\
&\leq \lambda t
\end{aligned}
$$

where at the second to last line we used that $\gamma \leq \lambda\frac{(1-\lambda)}{4}t_{\lambda/2}(\mathcal{Y}_n)$ (as $t_{\lambda/2}(\mathcal{Y}_n)$ is of order at least $n^{-1/d}$). This implies that $t_\lambda(\mathcal{X}_n) \leq t \leq t_{\lambda/2}(\mathcal{Y}_n)$. Using the upper bound on $t_{\lambda/2}(\mathcal{Y}_n)$ given in Theorem 4.6, we have that, with probability $1 - c(\ln n)^{\tilde{a}}n^{-b}$,

$$t^*(\mathcal{Y}_n) + \gamma \leq t_\lambda(\mathcal{X}_n) \leq \frac{2t^*(\mathcal{Y}_n)}{\lambda}\left(1 + C\left(\frac{(\ln n)^2}{n}\right)^{1/d}\right) \quad (43)$$

that is an analog of Theorem 4.6 also holds in a setting where tubular noise of size $(\ln n/n)^{2/d}$ is present.

6. Adaptive estimation with the selected scale

In this section, we show that the estimator $\hat{M} = \mathrm{Conv}(t_\lambda(\mathcal{X}_n), \mathcal{X}_n)$ is minimax adaptive on the scale of models $\mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}}$. For the sake of exposition, we focus on the noiseless case $\gamma = 0$. We first have to be careful when defining the scale of models. Indeed, by (9), we have for $\mu \in \mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}}$ supported on M

$$1 = \mu(M) \geq f_{\min}\omega_d\tau^d_{\min},$$

so that the model $\mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}}$ is empty if $f_{\min}\omega_d\tau^d_{\min} > 1$. Also, if we have $f_{\min}\omega_d\tau^d_{\min} = 1$, then μ is the uniform distribution on a sphere. In this case $d + 1$ observations characterize M, and the minimax rate on $\mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}}$ is zero for $n \geq d + 1$. To discard such degenerate cases, we will assume that there exists a constant $\kappa < 1$ so that $\omega_d f_{\min}\tau^d_{\min} < \kappa$. We have already mentioned in the introduction that Kim and Zhou [25] showed that the minimax risk $\mathcal{R}_n(\mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}})$ is of order $(\ln n/n)^{2/d}$. They were however not concerned

with precise constants. We indicate in Section B how to modify their proof to obtain a more precise result.

Proposition 6.1. *There exists a constant C depending only on κ such that*

$$\liminf_n \frac{\mathcal{R}_n(\mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}})}{(\ln n/n)^{2/d}} \geq \frac{C}{(\alpha_d f_{\min})^{2/d} \tau_{\min}}.$$

Our adaptivity result then reads as follows.

Theorem 6.2. *Let $d \geq 2$. Let $\mu \in \mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}}$ and let $0 < \lambda < (1+2/d)^{-1/d}$. Then, for n large enough, we have*

$$\mathbb{E}\left[d_H(\mathrm{Conv}(t_\lambda(\mathcal{X}_n), \mathcal{X}_n), M)\right] \leq \frac{c_0}{\lambda^2 (\alpha_d f_{\min})^{2/d} \tau_{\min}} \left(\frac{\ln n}{n}\right)^{2/d} \tag{44}$$
$$\leq \frac{c_1}{\lambda^2} \mathcal{R}_n(\mathcal{Q}^d_{\tau_{\min}}, f_{\min}, f_{\max}),$$

where c_0 is a numerical constant and c_1 only depends on κ.

Proof. Choose $b \in (0, 2]$ such that $\lambda < (1+b)^{-1/d} < (1+2/d)^{-1/d}$. Assume that the event described in (26) is satisfied (that is with probability larger than $1 - c(\ln n)^{d-1} n^{-b}$). Then, we have by Lemma 3.3

$$d_H(\mathrm{Conv}(t_\lambda(\mathcal{X}_n), \mathcal{X}_n), M) \leq \frac{t_\lambda(\mathcal{X}_n)^2}{\tau_{\min}} \leq \frac{t^*(\mathcal{X}_n)^2}{\lambda^2 \tau_{\min}} \left(1 + C\left(\frac{(\ln n)^2}{n}\right)^{1/d}\right)^2$$

We also assume that $\varepsilon(\mathcal{X}_n) \leq \left(\frac{4\ln n}{\alpha_d f_{\min} n n}\right)^{1/d}$, an event that happens with probability $1 - (\ln n)^{d-1} n^{-3}$ by Proposition 3.9. Then, for n large enough, we have $t^*(\mathcal{X}_n) \leq 2\varepsilon(\mathcal{X}_n)$ by Proposition 3.4. In particular, we obtain that, for n large enough

$$d_H(\mathrm{Conv}(t_\lambda(\mathcal{X}_n), \mathcal{X}_n), M) \leq \frac{c_0}{\lambda^2 (\alpha_d f_{\min})^{2/d} \tau_{\min}} \left(\frac{\ln n}{n}\right)^{2/d}$$

for some absolute constant c_0. The probability that this inequality is not satisfied is of order $(\ln n)^{d-1} n^{-b} \ll (\ln n/n)^{2/d}$, and if this is the case we bound the risk by $\mathrm{diam}(M)$ (that is bounded by a constant depending on $\tau_{\min}, f_{\min}$ and d [1, Lemma III.24]). We therefore obtain the first inequality of (44), while the second one follows directly from Proposition 6.1. □

Remark 6.3. *In the one-dimensional case $d = 1$, the minimax risk is of order $(\ln n/n)^2 / \left((\alpha_d f_{\min})^2 \tau_{\min}\right)$, whereas, with $b = 2$, the probability with which (43) holds is of order $(\ln n/n)^2$. As such, one can show that the risk of $\mathrm{Conv}(t_\lambda(\mathcal{X}_n), \mathcal{X}_n)$ is of order $(\ln n/n)^2$ for $d = 1$, but with a leading constant that will depend on the constants appearing in Theorem (4.6). This leading constant is therefore not anymore of order $1/\left((\alpha_d f_{\min})^2 \tau_{\min}\right)$, and we do not have*

a clean inequality of the form (44). *Still,* $\text{Conv}(t_\lambda(\mathcal{X}_n), \mathcal{X}_n)$ *is a data-driven minimax estimator even in this case.*

With a choice of λ smaller than $1/\sqrt{2}$ (say $\lambda = 1/2$), the condition $\lambda < (1 + 2/d)^{-1/d}$ is satisfied for every $d \geq 2$. With such a choice, we obtain a completely data-driven estimator that attains asymptotically the minimax rate $\mathcal{R}_n(\mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\max}})$ up to an absolute constant, for every admissible choice of $\tau_{\min}, f_{\min}, f_{\max}$ and $d \geq 2$. The slope λ in our selection procedure is akin to a regularization parameter that appears in most selection methods (such as in the LASSO [34], or the PCO and Goldenshluger-Lepski methods already mentioned). If every choice of parameter $\lambda < 1/\sqrt{2}$ is admissible from a theoretical point of view, the practical choice of the parameter λ is more delicate. We develop in Section 7 a heuristic, similar to the slope heuristics [9], to choose the parameter λ.

Remark 6.4. *We insist that our result is of an asymptotic nature, as the "large enough" in the above theorem depends on the probability measure μ. A similar behavior occurs with the PCO method mentioned in the introduction [27] (or with the Goldenshluger-Lepski method [26, Proposition 1]). Indeed, the remainder term $C(n, |\mathcal{H}|)$ appearing in* (5) *depends on μ through the ∞-norm of its density function, whereas the minimax risk does not depend on this ∞-norm (see [35, Theorem 2.8]). As such, the remainder term $C(n, |\mathcal{H}|)$ becomes negligible in front of the minimax risk only for n large enough with respect to μ (and not only with respect to the parameters defining the statistical model), as this is the case in Theorem* 6.2.

The parameter $t_\lambda(\mathcal{X}_n)$ actually gives us the approximation rate $\varepsilon(\mathcal{X}_n)$ up to a multiplicative constant (roughly equal to λ^{-1}). As such, it can be also used to design other data-driven estimators. As an example, we consider the estimation of the tangent spaces of a manifold. Let $x \in M$ and $A \subseteq M$ be a finite set. We denote by $T_x(A, t)$ the d-dimensional vector space U that minimizes $d_H(A \cap \mathcal{B}(x, t)|x + U)$. This estimator was originally studied in [12]. Recall that the angle between subspaces is denoted by $\angle$.

Corollary 6.5. *Let $\mu \in \mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{max}}$ with support M and let $0 < \lambda < (1 + 1/d)^{-1/d}$. Then, for n large enough (with respect to μ), we have*

$$\mathbb{E}\angle\left(T_x M, T_p\left(\mathcal{X}_n, 11 t_\lambda(\mathcal{X}_n)\right)\right) \leq c\left(\frac{\ln n}{n}\right)^{1/d}$$

for some constant c depending on $\lambda, d, \tau_{\min}$ and f_{min}

This rate is the minimax rate (up to logarithmic factors) according to [3, Theorem 3].

Proof. Theorem 3.2 in [12] states that for $A \subseteq M$, if $t < \tau(M)/2$ and $t \geq 10\varepsilon(A)$, then

$$\angle\left(T_x(A, t), T_x M\right) \leq 6\frac{t}{\tau(M)}.$$

As in the previous proof, we may choose $b \in (0, 2]$ such that $\lambda < (1 + b)^{-1/d} < (1 + 1/d)^{-1/d}$, and assume that the event described in Theorem 4.6 is satisfied. We also assume that $\varepsilon(\mathcal{X}_n) \leq \left(\frac{4 \ln n}{\alpha_d f_{\min} n}\right)^{1/d}$. Then, the quantity $t = 11 t_\lambda(\mathcal{X}_n)$ is larger than $10\varepsilon(\mathcal{X}_n)$ for n large enough, and furthermore satisfies $t \leq c_0 \left(\frac{\ln n}{\alpha_d f_{\min} n n}\right)^{1/d}$ for some absolute constant c_0 if n is large enough. We then have

$$\angle\left(T_x\left(\mathcal{X}_n, t\right), T_x M\right) \leq \frac{c_1}{\left(\alpha_d f_{\min}\right)^{1/d} \tau_{\min}} \left(\frac{\ln n}{n}\right)^{1/d}$$

for some absolute constant c_1 large enough. If one of the two conditions does not hold (this happens with a probability smaller than $(\ln n)^a n^{-b} = o\left(n^{-1/d}\right)$), we bound the angle by 2, concluding the proof. $\qquad\square$

Remark 6.6. *Authors in [14] also propose to use the convexity defect function of a set $A \subseteq M$ to estimate the reach of M, while their method requires only the knowledge of $\varepsilon(A)$. As such, we may use their technique by using the scale $t_\lambda(\mathcal{X}_n)$ instead of $\varepsilon(\mathcal{X}_n)$. This leads to a reach estimator that attains a risk of order $(\ln n/n)^{1/(3d)}$. As the minimax risk is of order $n^{-1/d}$ up to logarithmic factors for this problem (at least on a statistical model made of $\mathcal{C}^3$ manifolds), this is far from being minimax. Still, this yields a consistent fully data-driven reach estimator. We refer to [14] for details on the construction.*

7. Numerical considerations[1]

There are two distinct procedures to investigate: first, the computation of the t-convex hull $\mathrm{Conv}(t, \mathcal{X}_n)$, and second, the computation of the scale $t_\lambda(\mathcal{X}_n)$. To compute the t-convex hull $\mathrm{Conv}(t, \mathcal{X}_n)$, it suffices to compute the Čech complex $\mathrm{Cech}(t, \mathcal{X}_n) := \{\sigma \subseteq \mathcal{X}_n : r(\sigma) \leq t\}$. For $x \in \mathbb{R}^D$, let $N(x)$ be the number of points of $\mathcal{X}_n$ at distance less than $2t$ of x. Assume that one has access to the set $E_t(\mathcal{X}_n)$ of edges of $\mathcal{X}_n$ of length smaller than $2t$. Then, authors in [28] propose an algorithm of complexity $C_D \sum_{i=1}^n N\left(X_i\right)^D$ to compute $\mathrm{Cech}(t, \mathcal{X}_n)$. When t is of order $(\ln n/n)^{1/d}$, $N(X_i)$ is on average of order $\ln n$ and we obtain an average complexity of order $C_D n(\ln n)^D$. In high dimension, the complexity can be reduced if one has access to the dimension d by computing $\mathrm{Conv}_d(t, \mathcal{X}_n)$ instead (see Remark 3.8). Indeed, according to [28], the set of simplices of $\mathrm{Cech}(t, \mathcal{X}_n)$ of dimension smaller than d can be computed with average time complexity of order $C_d D n(\ln n)^d$. We also have to consider the computation of the edges $E_t(\mathcal{X}_n)$. A naive algorithm to compute this set leads to a complexity of order Dn^2, but in practice this can be considerably sped up by using e.g. a RP tree [17].

We now adress the selection procedure described in Section 4. To choose the scale $t_\lambda(\mathcal{X}_n)$, we have to compute the convexity defect function of $\mathcal{X}_n$. To do so, we need for each simplex $\sigma \subseteq \mathcal{X}_n$ to (i) compute its radius $r(\sigma)$ and (ii)

---

[1]Code is made available at github.com/vincentdivol/local-convex-hull.

compute $d_H(\mathrm{Conv}(\sigma)|\mathcal{X}_n)$. We will simplify this problem by considering only simplexes σ of dimension 1 (i.e. edges). Let $\mathrm{Graph}(t, \mathcal{X}_n)$ be the union of the edges of $\mathcal{X}_n$ of length $2t$. We may define a graph convexity defect function $\tilde{h}(t, \mathcal{X}_n) = d_H(\mathrm{Graph}(t, \mathcal{X}_n), \mathcal{X}_n)$, as well as a graph scale parameter

$$\tilde{t}_\lambda(\mathcal{X}_n) := \inf\left\{t \in \widetilde{\mathrm{Rad}}(\mathcal{X}_n) : \tilde{h}(t, \mathcal{X}_n) \leq \lambda t\right\},$$

where $\widetilde{\mathrm{Rad}}(\mathcal{X}_n) := \{|X_i - X_j|/2 : 1 \leq i, j \leq n\}$. A careful read of the proof of Theorem 4.6 shows that only edges are considered to obtain the different inequalities of the theorem. In particular, this theorem also holds with $\tilde{t}_\lambda(\mathcal{X}_n)$ instead of $t_\lambda(\mathcal{X}_n)$. When e is an edge of $\mathcal{X}_n$, the distance $d_H(\mathrm{Conv}(e)|\mathcal{X}_n)$ can be computed in $\mathcal{O}(n(D + \ln n))$ operations [7]. By looping over the $\mathcal{O}(n^2)$ edges of the dataset, we may compute $\tilde{h}(\cdot, \mathcal{X}_n)$ with a time complexity of $\mathcal{O}(n^3(D + \ln n))$.

The choice of the slope value λ has an impact on the selection procedure. Ideally, we would like to choose λ so that it is just below

$$\lambda_{\max}(\mathcal{X}_n) := \max\{\lambda : t_\lambda(\mathcal{X}_n) > t^*(\mathcal{X}_n)\}.$$

Let $t_{\max}(\mathcal{X}_n) = t_{\lambda_{\max}(\mathcal{X}_n)}(\mathcal{X}_n)$. According to Proposition 4.3, the function $h(\cdot, \mathcal{X}_n)$ is almost constant after $t^*(\mathcal{X}_n)$, and therefore also almost constant after $t_{\max}(\mathcal{X}_n)$. This implies that $t_\lambda(\mathcal{X}_n)$ should increase proportionally with $1/\lambda$ for $\lambda < \lambda_{\max}(\mathcal{X}_n)$ (at least approximately). On the opposite, for $\lambda > \lambda_{\max}(\mathcal{X}_n)$, we expect $t_\lambda(\mathcal{X}_n)$ to go to 0 quickly. By plotting the graph of the function $g_{\mathcal{X}_n} : \lambda \mapsto 1/t_\lambda(\mathcal{X}_n)$, those two behaviors should be observed (first linear and then diverging), so that a "jump" should occur around the value $\lambda_{\max}(\mathcal{X}_n)$. We indeed observe such a phenomenon, see Figure 8. In practice, we use a grid $0 = \lambda_1 \leq \cdots \leq \lambda_L = 1$ and the jump is defined by the smallest l such that the condition $g_{\mathcal{X}_n}(\lambda_{l+1}) - g_{\mathcal{X}_n}(\lambda_l) > 0.5 g_{\mathcal{X}_n}(0)$ is satisfied. We then select $\lambda_{\mathrm{choice}}(\mathcal{X}_n) = 0.8\lambda_{\mathrm{jump}}(\mathcal{X}_n)$ and let $t_{\mathrm{sel}}(\mathcal{X}_n) := t_{\lambda_{\mathrm{choice}}(\mathcal{X}_n)}(\mathcal{X}_n)$ (other constants than 0.5 and 0.8 would work as well).

Remark 7.1. *This method to select the slope λ is similar to the slope heuristics in model selection. Consider for instance the fixed-design regression setting where $Y = F + \varepsilon \in \mathbb{R}^n$ is observed with a Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathrm{Id})$. The goal is to reconstruct the signal F for the ℓ_2-loss, by selecting an estimator among the estimators $\hat{F}_m = \pi_{S_m}(Y)$, where $\{S_m : m\}$ is a collection of linear subspaces, each S_m being of dimension D_m. A classical method to select the estimator F_m is to choose*

$$\hat{m}(C) \in \underset{m}{\arg\min}\left\{\left|\hat{F}_m - Y\right|^2 + CD_m\right\}$$

where C is a constant to fix. In theory, any value of C smaller than σ^2 will lead to overfitting, whereas values of C larger than σ^2 are admissible. We then say that $C = \sigma^2$ is the minimal penalty. The exact value of the minimal penalty $C = \sigma^2$ is of an asymptotic nature. However, we still see a minimal penalty phenomenon occurring in practice: for C too small, the selected dimension $D_{\hat{m}(C)}$ will be very
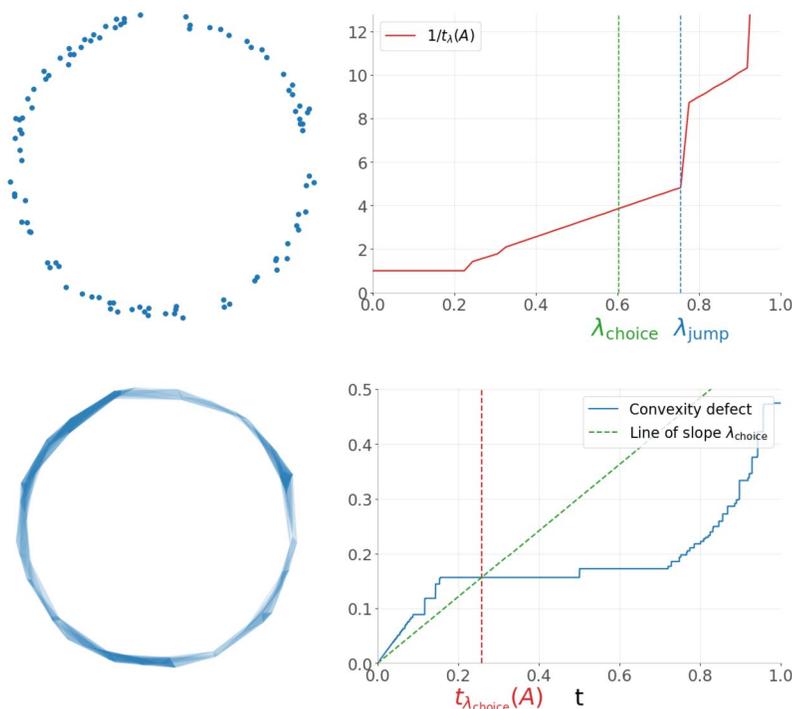
FIG 8. *Top left. Sample $\mathcal{X}_n$. Top right. The value of $\lambda_{choice}(\mathcal{X}_n)$ is equal to $0.8\lambda_{jump}(\mathcal{X}_n)$. Bottom left. The set $\mathrm{Conv}(t_{sel}(\mathcal{X}_n), \mathcal{X}_n)$. Bottom right. The graph convexity defect function $h(\cdot, \mathcal{X}_n)$.*

large, whereas at some value $\hat{C}_{jump}$ it will suddenly decrease and gets smaller. This jump is detected and is used to select the value of C. We refer to [9] for details. A similar phenomenon occurs in our setting: the slope λ plays the role of the parameter C (or rather $1/C$), and we have a maximal penalty phenomenon: every value of λ smaller than 1 is theoretically admissible. The quantity $1/t$ is the analog of the dimension D_m, as it is a measure of the complexity of the estimator $\mathrm{Conv}(t, \mathcal{X}_n)$: choosing $t = +\infty$ amounts to assuming that M is a convex set, whereas choosing very small values of t amounts to assuming that M has a small reach. In practice, we observe a jump in the function $g_{\mathcal{X}_n} : \lambda \mapsto 1/t_\lambda(\mathcal{X}_n)$, and we use this phenomenon to choose the parameter λ.

In Figure 8, we display the graph convexity defect function $\tilde{h}(\cdot, \mathcal{X}_n)$ for a set $\mathcal{X}_n$ made of $n = 100$ uniformly sampled point on the unit circle M, with a tubular uniform noise of size $\gamma = 0.1$. Both the "jump" phenomenon in the function $g_{\mathcal{X}_n}$ and the expected behavior of the function $h(\cdot, \mathcal{X}_n)$ occur. We evaluate $\varepsilon(\mathcal{X}_n) = 0.16$, while $\lambda_{\mathrm{choice}}(\mathcal{X}_n) = 0.60$ and $t_{\mathrm{sel}}(\mathcal{X}_n) = 0.26$. According to [31, Proposition 3.1], the Cech complex $\mathrm{Cech}(\mathcal{X}_n, 2t)$ on A of radius $2t$ has the same homology as M as long as $t \geq \varepsilon(\mathcal{X}_n)$. As a safety check, we compute the
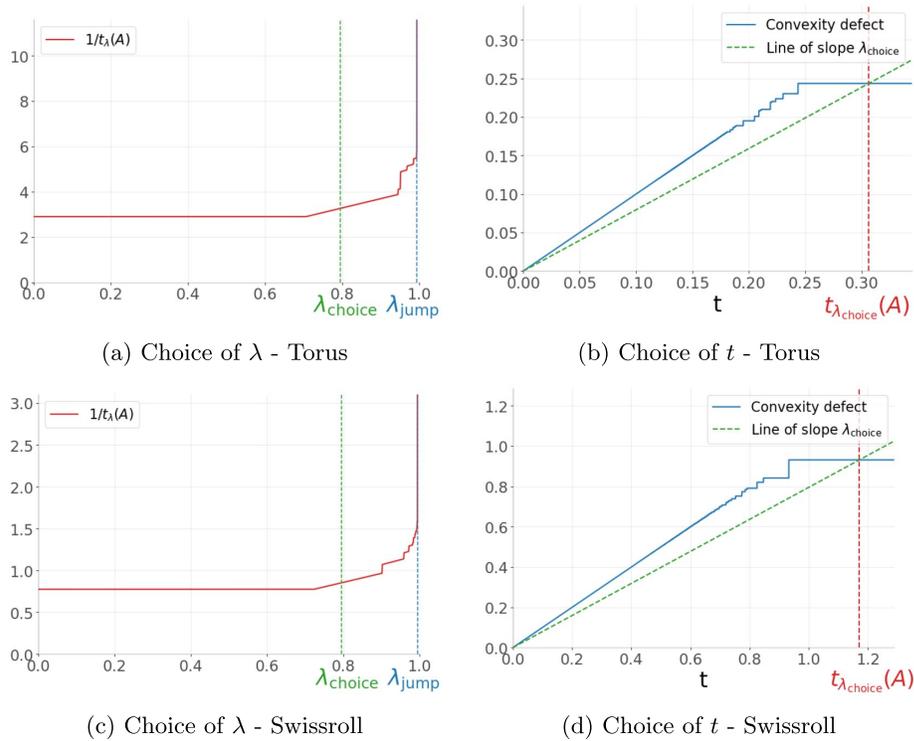
(a) Choice of λ - Torus

(b) Choice of t - Torus

(c) Choice of λ - Swissroll

(d) Choice of t - Swissroll

FIG 9. *For a set $\mathcal{X}_n$ made of 10^4 points sampled on the torus (resp. on the swiss roll), we compute $g_{\mathcal{X}_n}$ and $h(\cdot, \mathcal{X}_n)$ up to the value $t = \ell_{K_{\max}}(\mathcal{X}_n)$. The selected values of λ are respectively 0.796 and 0.792, while the selected values of $t_{sel}(\mathcal{X}_n)$ are 0.309 and 1.126. In both cases, we also estimate the approximation rate $\varepsilon(\mathcal{X}_n)$, respectively equal to 0.254 and 0.891. Both times, we indeed have $t_{sel}(\mathcal{X}_n) \geq \varepsilon(\mathcal{X}_n)$, and furthermore, the Čech complex of parameter $2t_{sel}(\mathcal{X}_n)$ has the same homology as the torus (resp. the swiss roll).*

homology of $\mathrm{Cech}\,(\mathcal{X}_n, 2t_{\mathrm{sel}}(\mathcal{X}_n))$, which is indeed equal to the homology of the circle.

Actually, it is not necessary to compute the whole convexity defect function to compute $\tilde{t}_\lambda(\mathcal{X}_n)$, as one can stop at the first value for which $\tilde{h}(t, \mathcal{X}_n) < \lambda t$. This can be used to speed up the computation $t_{\mathrm{sel}}(\mathcal{X}_n)$. Given an integer K, we let $\ell_K(\mathcal{X}_n)$ be half the maximum distance between a point of $\mathcal{X}_n$ and its K th nearest neighbor in $\mathcal{X}_n$. We compute for each point X_i in $\mathcal{X}_n$ its K nearest neighbors $\mathcal{X}_K^i$ (using for instance a RP tree [17]). Then, for each point X_j in $\mathcal{X}_K^i$, if $e = (X_i, X_j)$, we have $d_H(\mathrm{Conv}(e)|\mathcal{X}_n) = d_H(\mathrm{Conv}(e)|\mathcal{X}_K^i)$. The latter distance can be computed in $\mathcal{O}(K(D + \ln K))$ operations. There are at most nK such edges, so that we compute $\tilde{h}(\cdot, \mathcal{X}_n)$ up to $t = \ell_K(\mathcal{X}_n)$ with $\mathcal{O}(nK^2(D + \ln K))$ operations. We then apply the slope selection procedure on the convexity defect function up to $\ell_K(\mathcal{X}_n)$. If we select the maximal value possible, that is if $t_{\mathrm{sel}}(\mathcal{X}_n) = \ell_K(\mathcal{X}_n)$, then we did not go far enough in the computation of the
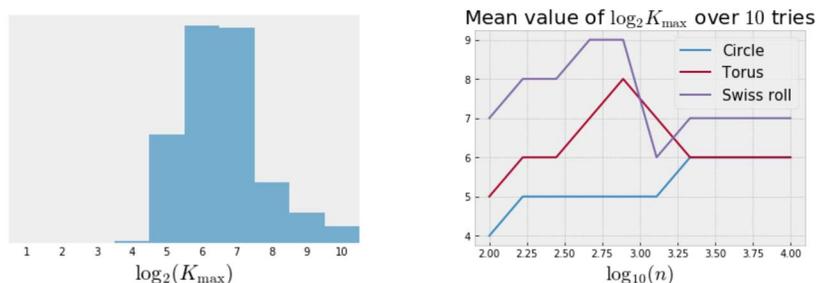
FIG 10. *Left. Distribution of* $\log_2 K_{\max}$ *over the different point clouds (circle, torus and swiss roll of different sizes on 10 tries each). Right. For each class and each number of points* n, *we display the mean value of* $\log_2 K_{\max}$ *over the 10 tries: in each class, it stays bounded as* n *grows. Large values of* $K_{\max}$ *for the swiss roll dataset correspond to numbers of samples* n *for which* $\varepsilon(\mathcal{X}_n)$ *is too large* $(n \leq 1000)$: *the subquadratic behavior then does not occur and therefore the whole convexity defect function is computed.*

convexity defect function. In that case, we repeat the procedure with $\bar{K} = 2K$. If $t_{\mathrm{sel}}(\mathcal{X}_n) < \ell_K(\mathcal{X}_n)$, we stop. In practice, the maximal value $K_{\max}$ of K is much smaller than n and this approach leads to a considerable speed-up.

We test this faster algorithm on three classes of datasets. The first class is made of n points uniformly sampled on a circle that lies on a random plane in $\mathbb{R}^{100}$, that are corrupted with uniform noise (in $\mathbb{R}^{100}$) of size $(\ln n/n)^{2/d}$. The second class consists of points sampled on the torus of inner radius 1 and outer radius 4. The third class is made of points sampled on the swiss roll dataset from the SciPy Python library [36]. For each class, we conduct 10 experiments for each value of n, n ranging from 10^2 to 10^4. The value $K_{\max}$ was never larger than $2^{10} = 1024$, and did not increase with n, see Figure 10. Increasing the ambient dimension in the first class did not significantly increase the computation time. We display in Figure 9 the functions $\hat{h}(\cdot, \mathcal{X}_n)$ and $g_{\mathcal{X}_n}$ for two point clouds from this dataset: we observe once again the "jump" phenomenon occurring.

8. Discussion and further works

In this article, we introduced a particularly simple manifold estimator, based on a unique rule: add the convex hull of any subset of the set of observations which is of radius smaller than t. After proving that this leads to a minimax estimator for some choice of t, we explained how to select the parameter t by computing the convexity defect function of the set of observations. The selection procedure actually allows us to find a parameter $t_\lambda(\mathcal{X}_n)$ such that $\varepsilon(\mathcal{X}_n)/t_\lambda(\mathcal{X}_n)$ is arbitrarily close to 1 (by choosing λ close enough to 1). The selected parameter can therefore be used as a scale parameter in a wide range of procedures in geometric inference. We illustrated this general idea by showing how a data-driven minimax tangent space estimator can be created thanks to $t_\lambda(\mathcal{X}_n)$. The main limitation to our procedure is its non-robustness to outliers. Indeed, even in the presence of one outlier in $\mathcal{X}_n$, the loss function $t \mapsto d_H(\mathrm{Conv}(t, \mathcal{X}_n), M)$

would be constant, equal to the distance between the outlier and the manifold M: with respect to the Hausdorff distance, all the estimators $\mathrm{Conv}(t, \mathcal{X}_n)$ are then equally bad. Of course, even in that case, we would like to assert that some values of t are "better" than others in some sense. A solution to overcome this issue would be to change the loss function, for instance by using Wasserstein distances on judicious probability measures built on the t-convex hulls $\mathrm{Conv}(t, \mathcal{X}_n)$ in place of the Hausdorff distance

Appendix A: Proof of Lemma 4.12

Let $S = \sum_{k=1}^{K} \mathbf{1}\{N_k = 2\}$. Let $\tilde{n}$ be the number of points of $\mathcal{X}_n$ in $\bigcup_k U_k$, so that $\tilde{n}$ follows a binomial distribution of parameters n and Km. Recall that by construction, $Km \geq c_0$ for some constant c_0 (see Lemma 4.9). Conditionally on $\tilde{n}$, the random variable S can be realized as the number of urns containing exactly two balls, in a model where $\tilde{n}$ balls are thrown uniformly in K urns. Let $p_i = \binom{\tilde{n}}{i} K^{-i} \left(1 - K^{-1}\right)^{\tilde{n}-i}$ be the probability that an urn contains exactly i balls. We have $\mathbb{E}[S|\tilde{n}] = Kp_2$, and

$$\mathbb{E}\left[\exp\left(-C_1 S\right)|\tilde{n}\right] \leq \mathbb{E}\left[\exp\left(-C_1 Kp_2/2\right)\mathbf{1}\{S \geq Kp_2/2\}|\tilde{n}\right] + \mathbb{P}\left(S < Kp_2/2|\tilde{n}\right)$$
$$\leq \exp\left(-C_1 Kp_2/2\right) + \mathbb{P}\left(|S - Kp_2| > Kp_2/2|\tilde{n}\right).$$
(45)

Let $v = 2K \max\left(2p_2, 3p_3\right)$. According to [13, Proposition 3.5], if for some $s > 0$,

$$Kp_2/2 \geq \sqrt{4vs} + 2s/3,$$
(46)

then $\mathbb{P}\left(|S - Kp_2| > Kp_2/2|\tilde{n}\right) \leq 4e^{-s}$. Recall that $nm^2 \leq 1$ by assumption, and that $K \geq c_{\mu,\delta} t^{-d} \geq c_1/m$. We therefore have $n/K^2 \leq c_1^{-2}$. Assuming that $\tilde{n} \geq 3$ and using the inequality $\ln\left(1 - K^{-1}\right) \geq -K^{-1} - K^{-2}$ for $K \geq 2$, we obtain the inequalities

$$p_2 \geq \frac{(\tilde{n}/K)^2}{4e^{c_1^{-2}}} e^{-\tilde{n}/K} \text{ and } p_3 \leq \frac{e^3}{6}(\tilde{n}/K)^3 e^{-\tilde{n}/K} \leq c_2 p_2(n/K)$$
(47)

for some positive constant c_2. We consider two different regimes.

- Assume first that $n/K \leq 2/\left(3c_2\right)$. Then $3p_3 \leq 2p_2$ and one can check that $s = Kp_2/100$ satisfies (46). Inequality (45) then yields that

$$\mathbb{E}\left[\exp\left(-C_1 S\right)|\tilde{n}\right] \leq 5\exp\left(-C_1' Kp_2\right)$$

for $C_1' = \min\left(C_1/2, 1/100\right)$. To conclude, we remark that for any $\alpha \in (0,1)$, by the Hoeffding inequality, the event $|\tilde{n} - nKm| \leq nKm\alpha$ holds with probability at least $1 - \exp\left(-2n\alpha^2\right)$. Letting $\alpha = 1/2$, we obtain that, on this event,

$$\frac{1}{2}nm \leq \frac{\tilde{n}}{K} \leq \frac{3}{2}nm \leq \frac{3}{2}\frac{n}{K}mK \leq \frac{1}{c_2}$$

where we used that $mK \leq 1$. Therefore, $p_2 \geq c_3(nm)^2 \geq c_4(nm)^2 e^{-nm}$ for some constants c_3 and c_4. The probability of order $\exp\left(-2n\alpha^2\right)$ being negligible, we obtain a final bound of order $\exp\left(-C_1' c_4 K(nm)^2 e^{-nm}\right) \leq \exp\left(-C_2 n\phi(nm)\right)$, concluding the proof in the regime $n/K \leq 2/(3c_2)$.

- Otherwise, we have $n/K > 2/(3c_2)$ and we also assume that $|\tilde{n} - nKm| \leq \alpha nKm$ for some $\alpha \in (0,1)$ to fix (this happens with probability $1 - \exp\left(-2n\alpha^2\right)$ by Hoeffding's inequality). One can then check using (47) that $s = c_5 \tilde{n} e^{-\tilde{n}/K}$ satisfies (46) if c_5 is chosen small enough. Furthermore, $s \leq c_6 K p_2$ for some constant c_6 (using (47)). The leading term in (45) is therefore of the form $\exp\left(-c_7 \tilde{n} e^{-n/K}\right)$. Let $\alpha = 1/(\ln n)^3$. We have, as $nm \geq c_0 n/K \geq c_8$ and as $nm \leq (\ln n)^2$ (by assumption),

$$c_9 \leq nm(1-\alpha) \leq \frac{\tilde{n}}{K} \leq nm(1+\alpha) \leq nm + \frac{1}{\ln n}.$$

Therefore, $\tilde{n} e^{-\tilde{n}/K} \geq (c_9/2) K e^{-nm}$. The probability of order $\exp(-2n\alpha^2)$ is still negligible, and we obtain a final bound on $\mathbb{E}\left[\exp\left(-C_1 S\right)\right]$ of order $\exp\left(-(c_9/2) K e^{-nm}\right) \leq \exp\left(-c_{10} n\phi(nm)\right)$.

Appendix B: Precise lower bound on the minimax risk

We adapt the construction made in [25] so that the lower bound on the minimax risk holds with an explicit constant. Let $0 < d < D$ and $\tau_{\min}, f_{\min}, f_{\max}$ with $\omega_d f_{\min} \tau_{\min}^d < \kappa$. We let $M(\mu)$ be the underlying manifold of the law $\mu \in \mathcal{Q}_{\tau_{\min}, f_{\min}, f_{\max}}^d$. The lowerbound is based on Le Cam's lemma:

Lemma B.1. *Let $\mathcal{P}^{(1)}, \mathcal{P}^{(2)}$ be two subfamilies of $\mathcal{Q}_{\tau_{\min}, f_{\min}, f_{\max}}^d$ which are ε-separated, in the sense that $d_H(M(\mu^{(1)}), M(\mu^{(2)})) \geq 2\varepsilon$ for all $\mu^{(1)} \in \mathcal{P}^{(1)}, \mu^{(2)} \in \mathcal{P}^{(2)}$. Then*

$$\mathcal{R}_n(M, \mathcal{Q}_{\tau_{\min}, f_{\min}, f_{\max}}^d) \geq \varepsilon \left| \left(\frac{1}{\#\mathcal{P}^{(1)}} \sum_{\mu^{(1)} \in \mathcal{P}^{(1)}} \mu^{(1)} \right) \wedge \left(\frac{1}{\#\mathcal{P}^{(2)}} \sum_{\mu^{(2)} \in \mathcal{P}^{(2)}} \mu^{(2)} \right) \right|,$$

(48)

where $|\mu \wedge \nu|$ is the testing affinity between two distributions μ and ν.

To obtain a lowerbound on the minimax risk, authors in [25] exhibit two families of manifolds which are ε-separated, and consider the uniform distributions on them. Those manifolds are built by considering a base manifold M_0 which is locally flat, and by adding small bumps on the locally flat part. Such a construction leads to distributions having a density equal roughly to $1/\operatorname{Vol}(M_0)$, a constant which might be smaller than $f_{\min}$. If this is the case, then the corresponding submodels are not in $\mathcal{Q}_{\tau_{\min}, f_{\min}, f_{\max}}^d$ and we cannot apply Le Cam's Lemma. Hence, we consider another base manifold, which is a sphere M_0 of radius R slightly larger than $\tau_{\min}$, so that its volume is smaller than $1/f_{\min}$ (this is possible as $f_{\min} \omega_d \tau_{\min}^d \leq \kappa < 1$). The two families are then once again constructed by adding small bumps on M_0. We now detail this construction.

Let $R, \delta > 0$ be two parameters to be fixed later. Let $M_0 \subseteq \mathbb{R}^{d+1} \subseteq \mathbb{R}^D$ be the d-sphere of radius R, and let A be a maximal subset of M_0 of even size, which is 4δ-separated. Note that, standard packing arguments (and the formula for the volume of a spherical cap) show that if δ/R is small enough, then the cardinality $2m$ of A satisfies $2m \geq \left(\frac{c_0 R}{\delta}\right)^d$ for some absolute constant c_0.

Let $\phi : \mathbb{R} \to \mathbb{R}$ be a smooth function such that $0 \leq \phi \leq 1, \phi \equiv 1$ on $[-1, 1]$ and $\phi \equiv 0$ on $\mathbb{R}\backslash[-2, 2]$. For $s \in \{\pm 1\}^A$, we build a diffeomorphism Φ_s^ε by letting for $x \in \mathbb{R}^D$

$$\Phi_s^\varepsilon(x) = x\left(1 + \frac{\varepsilon}{R}\sum_{y \in A} s(y)\phi\left(\frac{\|x - y\|}{\delta}\right)\right). \tag{49}$$

Recall that $\|N\|_{\mathrm{op}}$ denotes the operator norm of a linear application N.

Lemma B.2. *There exists two absolute constants $c_0, c_1, c_2 > 0$ such that the following holds. Assume that $\delta \leq R$ and that $c_0\varepsilon/\delta < 1$. Then, the function $\Phi_s^\varepsilon : \mathcal{B}(0, 3R) \to \mathbb{R}^{d+1}$ is a diffeomorphism on its image, with*

$$\sup_{x \in \mathcal{B}(0,3R)} \|\mathrm{Id} - d_x\Phi_s^\varepsilon\|_{\mathrm{op}} \leq c_1\varepsilon/\delta \text{ and } \sup_{x \in \mathcal{B}(0,3R)} \|d_x^2\Phi_s^\varepsilon\|_{op} \leq c_2\varepsilon/\delta^2. \tag{50}$$

Proof. As A is 4δ-separated, at most one term in the sum in (B.2) is non-zero. A computation gives that the derivative of Φ_B is given by, for $x \in \mathcal{B}(0, 3R)$,

$$d_x\Phi_s^\varepsilon(h) =$$
$$h + h\frac{\varepsilon}{R}\sum_{y \in A} s(y)\phi\left(\frac{|x - y|}{\delta}\right) + x\frac{\varepsilon}{R}\sum_{y \in A}\frac{1}{\delta}s(y)\phi'\left(\frac{|x - y|}{\delta}\right)\frac{\langle x - y, h\rangle}{|x - y|}. \tag{51}$$

Hence,

$$\|\mathrm{Id} - d_x\Phi_s^\varepsilon\|_{\mathrm{op}} \leq \frac{\varepsilon}{R}\left(\|\phi\|_\infty + |x|\frac{\|\phi'\|_\infty}{\delta}\right) \leq \frac{\varepsilon}{R}\left(\|\phi\|_\infty + 3R\frac{\|\phi'\|_\infty}{\delta}\right) \leq c_1\frac{\varepsilon}{\delta}$$

where $c_1 = c_0\|\phi\|_\infty + 3\|\phi'\|_\infty$. A similar computation gives that $\|d_x^2\Phi_s^\varepsilon\|_{\mathrm{op}} \leq c_2\varepsilon/\delta^2$ for $c_2 = 4\|\phi'\|_\infty + 3\|\phi''\|_\infty$. We eventually show the injectivity: if $\Phi_s^\varepsilon(x) = \Phi_s^\varepsilon(x')$, then x and x' are colinear. Also, if $c_0 = \|\phi\|_\infty + 3\|\phi'\|_\infty$, one can check using (51) that the derivative of the function $r \in [0, 3R] \mapsto \langle\Phi_s^\varepsilon(ru), u\rangle$ for u an unit vector is increasing, proving the injectivity. $\square$

Therefore, from [19, Theorem 14.19], we infer that $M_s^\varepsilon := \Phi_s^\varepsilon(M)$ is a manifold with reach larger than

$$\tau(M_s^\varepsilon) \geq R\min\left(1 - c_1\varepsilon/\delta, \frac{(1 - c_1\varepsilon/\delta)^2}{1 + c_1\varepsilon/\delta + Rc_2\varepsilon/\delta^2}\right) \tag{52}$$

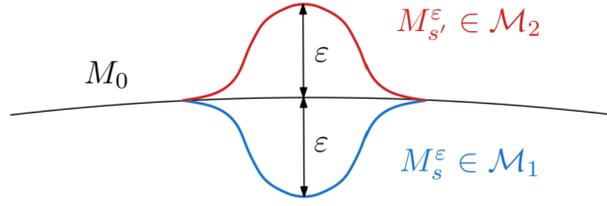FIG 11. *An element $\mu^{(1)} \in \mathcal{P}^{(1)}$ has its first marginal supported on the blue manifold M_s^ε (lower bump), whereas an element $\mu^{(2)} \in \mathcal{P}^{(2)}$ is supported on the red manifold $M_{s'}^\varepsilon$ (upper bump).*

Denote by $J\Phi_s^\varepsilon$ the Jacobian of Φ_ε^s. Then, the volume of M_s^ε is controlled by

$$\omega_d R^d \leq \mathrm{Vol}\,(M_s^\varepsilon) = \int_{M_0} J\Phi_s^\varepsilon(x)\mathrm{d}x = \omega_d R^d + \sum_{y \in A} \int_{\mathcal{B}_{M_0}(y,2\delta)} (J\Phi_s^\varepsilon(x) - 1)\,\mathrm{d}x$$

$$\leq \omega_d R^d + 2m C_d c_1 \frac{\varepsilon}{\delta}\,\mathrm{Vol}\,(\mathcal{B}_{M_0}(y,2\delta)) \leq \omega_d R^d \left(1 + C_d c_1 \frac{\varepsilon}{\delta}\right) \tag{53}$$

where we used that $\det(N) - 1 \leq C_d\|N - \mathrm{Id}\|_{\mathrm{op}}$ for some constant C_d if N is a matrix of size d with operator norm smaller than 1, the fact that $2m\,\mathrm{Vol}\,(\mathcal{B}_{M_0}(y,2\delta)) \leq \mathrm{Vol}\,(M_0)$, and Lemma B.2.

Let $R = \tau_{\min} + \frac{1}{2}\left(\frac{1}{(\omega_d f_{\min})^{1/d}} - \tau_{\min}\right)$ and $\delta = \sqrt{R\varepsilon}\nu$ where $\nu^2 = \frac{2c_2\tau_{\min}}{R - \tau_{\min}}$. With this choice of parameters, one can check that, for ε/δ small enough, $\tau\,(M_s^\varepsilon) \geq \tau_{\min}$ (by (52)) and $\mathrm{Vol}\,(M_s^\varepsilon) \leq 1/f_{\min}$ (by (53) and using that $\omega_d f_{\min}\tau_{\min}^d \leq \kappa < 1$).

We define the family $\mathcal{M}^{(1)}$ of manifolds M_s^ε where s contains exactly m signs $+1$ (and m signs -1). The family $\mathcal{M}^{(2)}$ is defined likewise by considering M_s^ε where s contains exactly $m + 1$ or $m - 1$ signs $+1$. We then let $\mathcal{P}^{(1)}$ be the set of distributions Q_s^ε where Q_s^ε is the uniform distribution on a manifold of $M_s^\varepsilon \in \mathcal{M}^{(1)}$, so that $\mathcal{P}^{(1)}$ is a subset of $\mathcal{Q}_{\tau_{\min}, f_{\min}}^d$ fhe set $\mathcal{P}^{(2)}$ is defined likewise.

By construction, the two families $\mathcal{P}^{(1)}, \mathcal{P}^{(2)}$ are 2ε-separated (see Figure 11). Hence, we can apply Le Cam's lemma. The exact same computations than in [25, Section 3] show that the testing affinity between $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ converge to 1 as long as $4m = n/\ln n$. Thus, Le Cam's Lemma (48) yields

$$\liminf_n \frac{\mathcal{R}_n(M, \mathcal{Q}_{\tau_{\min}, f_{\min}, f_{\max}}^d)}{(\ln n/n)^{2/d}} \geq \liminf_n (m/4)^{2/d}\varepsilon. \tag{54}$$

As $2m \geq (c_0 R/\delta)^d$, we therefore have

$$\liminf_n \frac{\mathcal{R}_n(M, \mathcal{Q}_{\tau_{\min}, f_{\min}, f_{\max}}^d)}{(\ln n/n)^{2/d}} \geq \frac{c_0^2}{8^{2/d}} \frac{R^2}{\delta^2}\varepsilon = \frac{c_0^2}{8^{2/d}} \frac{R}{\nu^2}$$

$$= \frac{c_0^2}{8^{2/d}} \frac{R\,(R - \tau_{\min})}{2c_2\tau_{\min}} \geq \frac{c_3}{(\omega_d f_{\min})^{1/d}\tau_{\min}} \left(\frac{1}{(\omega_d f_{\min})^{1/d}} - \tau_{\min}\right),$$

for some absolute constant c_3, where we used that by definition,

$$R - \tau_{\min} = \frac{1}{2}\left(\frac{1}{(\omega_d f_{\min})^{1/d}} - \tau_{\min}\right),$$

and that $R \geq \frac{1}{2}\left(\omega_d f_{\min}\right)^{-1/d}$. As $\tau_{\min} \leq \kappa/\left(\omega_d f_{\min}\right)^{1/d}$, and as $\omega_d^{1/d} \leq c\alpha_d^{1/d}$ for some absolute constant c, we obtain the conclusion with constant $C = c_3(1 - \kappa)/c$. Note that the lower bound actually holds on the smaller model $\mathcal{Q}^d_{\tau_{\min}, f_{\min}, f_{\min}}$, as we only considered uniform distributions in the proof.

Acknowledgments

I am grateful to Fréderic Chazal (Inria Saclay) and Pascal Massart (Université Paris-Sud) for thoughtful discussions and valuable comments on both mathematical and computational aspects of this work. I would also like to thank the anonymous reviewers for their helpful suggestions.

References

[1] AAMARI, E. (2017). Vitesses de convergence en inférence géométrique, PhD thesis, Paris Saclay.
[2] AAMARI, E. and LEVRARD, C. (2018). Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction. *Discrete & Computational Geometry* **59** 923–971. MR3802310
[3] AAMARI, E. and LEVRARD, C. (2019). Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics* **47** 177–204. MR3909931
[4] AAMARI, E., KIM, J., CHAZAL, F., MICHEL, B., RINALDO, A. and WASSERMAN, L. (2019). Estimating the Reach of a Manifold. *Electronic Journal of Statistics* **13** 1359–1399. MR3938326
[5] AARON, C. and BODART, O. (2016). Local convex hull support and boundary estimation. *Journal of Multivariate Analysis* **147** 82–101. MR3484171
[6] ALMGREN, F. (1986). Optimal isoperimetric inequalities. *Indiana University mathematics journal* **35** 451–547. MR0855173
[7] ALT, H., BRASS, P., GODAU, M., KNAUER, C. and WENK, C. (2003). Computing the Hausdorff distance of geometric patterns and shapes. In *Discrete and computational geometry* 65–76. Springer. MR2038470
[8] ARIAS-CASTRO, E. and LE GOUIC, T. (2019). Unconstrained and curvature-constrained shortest-path distances and their approximation. *Discrete & Computational Geometry* **62** 1–28. MR3959920
[9] ARLOT, S. (2019). Minimal penalties and the slope heuristics: a survey. *Journal de la société française de statistique* **160** 1–106. MR4021408
[10] ATTALI, D., LIEUTIER, A. and SALINAS, D. (2013). Vietoris–Rips complexes also provide topologically correct reconstructions of sampled shapes. *Computational Geometry* **46** 448–465. MR3003901

[11] Balakrishnan, S., Rinaldo, A., Sheehy, D., Singh, A. and Wasserman, L. (2012). Minimax rates for homology inference. In *Artificial Intelligence and Statistics* 64–72.

[12] Belkin, M., Sun, J. and Wang, Y. (2009). Constructing Laplace operator from point clouds in $\mathbb{R}^d$. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms* 1031–1040. Society for Industrial and Applied Mathematics. MR2807545

[13] Ben-Hamou, A., Boucheron, S. and Ohannessian, M. I. (2017). Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli* **23** 249–287. MR3556773

[14] Berenfeld, C., Harvey, J., Hoffmann, M. and Shankar, K. (2021). Estimating the reach of a manifold via its convexity defect function. *Discrete & Computational Geometry* 1–36.

[15] Boissonnat, J.-D. and Ghosh, A. (2014). Manifold reconstruction using tangential Delaunay complexes. *Discrete & Computational Geometry* **51** 221–267. MR3148657

[16] Cheng, S.-W. and Chiu, M.-K. (2016). Tangent estimation from point samples. *Discrete & Computational Geometry* **56** 505–557. MR3544007

[17] Dasgupta, S. and Freund, Y. (2008). Random projection trees and low dimensional manifolds. In *Proceedings of the fortieth annual ACM symposium on Theory of computing* 537–546. MR2582911

[18] David, H. A. and David, H. A. (1970). *Order Statistics. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series.* Wiley. MR0597893

[19] Federer, H. (1959). Curvature measures. *Transactions of the American Mathematical Society* **93** 418–491. MR0110078

[20] Federer, H. (1969). *Geometric measure theory. Grundlehren der mathematischen Wissenschaften.* Springer. MR0257325

[21] Genovese, C. R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2012a). Manifold estimation and singular deconvolution under Hausdorff loss. *The Annals of Statistics* **40** 941–963. MR2985939

[22] Genovese, C. R., Perone Pacifico, M., Isabella, V. and Wasserman, L. (2012b). Minimax manifold estimation. MR2930639

[23] Hein, M. and Audibert, J.-Y. (2005). Intrinsic dimensionality estimation of submanifolds in $\mathbb{R}^d$. In *Proceedings of the 22nd international conference on Machine learning* 289–296. ACM.

[24] Kim, J., Rinaldo, A. and Wasserman, L. (2019). Minimax rates for estimating the dimension of a manifold. *Journal of Computational Geometry* **10** 42–95. MR3918925

[25] Kim, A. K. H. and Zhou, H. H. (2015). Tight minimax rates for manifold estimation under Hausdorff loss. *Electronic Journal of Statistics* **9** 1562–1582. MR3376117

[26] Lacour, C. and Massart, P. (2016). Minimal penalty for Goldenshluger–Lepski method. *Stochastic Processes and their Applications* **126** 3774–3789. MR3565477

[27] Lacour, C., Massart, P. and Rivoirard, V. (2017). Estimator selec-

tion: a new method with applications to kernel density estimation. *Sankhya A* **79** 298–335. MR3707423

[28] LE, N.-K., MARTINS, P., DECREUSEFOND, L. and VERGNE, A. (2015). Construction of the generalized czech complex. In *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)* 1–5. IEEE.

[29] LITTLE, A. V., JUNG, Y.-M. and MAGGIONI, M. (2009). Multiscale estimation of intrinsic dimensionality of data sets. In *2009 AAAI Fall Symposium Series*.

[30] MAGGIONI, M., MINSKER, S. and STRAWN, N. (2016). Multiscale dictionary learning: non-asymptotic bounds and robustness. *The Journal of Machine Learning Research* **17** 43–93. MR3482922

[31] NIYOGI, P., SMALE, S. and WEINBERGER, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry* **39** 419–441. MR2383768

[32] PUCHKIN, N. and SPOKOINY, V. (2019). Structure-adaptive manifold estimation. *arXiv preprint arXiv:1906.05014*.

[33] RODRÍGUEZ CASAL, A. (2007). Set estimation under convexity type assumptions. In *Annales de l'IHP Probabilités et statistiques* **43** 763–774. MR3252430

[34] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58** 267–288. MR1379242

[35] TSYBAKOV, A. (2008). *Introduction to Nonparametric Estimation. Springer Series in Statistics.* Springer New York. MR2724359

[36] VIRTANEN, P., GOMMERS, R., OLIPHANT, T. E., HABERLAND, M., REDDY, T., COURNAPEAU, D., BUROVSKI, E., PETERSON, P., WECKESSER, W., BRIGHT, J., VAN DER WALT, S. J., BRETT, M., WILSON, J., JARROD MILLMAN, K., MAYOROV, N., NELSON, A. R. J., JONES, E., KERN, R., LARSON, E., CAREY, C., POLAT, İ., FENG, Y., MOORE, E. W., VAND ERPLAS, J., LAXALDE, D., PERKTOLD, J., CIMRMAN, R., HENRIKSEN, I., QUINTERO, E. A., HARRIS, C. R., ARCHIBALD, A. M., RIBEIRO, A. H., PEDREGOSA, F., VAN MULBREGT, P. and CONTRIBUTORS, S. . . (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*.

[37] ZHANG, Y. and DING, X. (2017). Limit properties for ratios of order statistics from exponentials. *Journal of inequalities and applications* **2017** 1–8. MR3592885