Minimax adaptive estimation in manifold inference

Vincent Divol

MINIMAX ADAPTIVE ESTIMATION IN MANIFOLD INFERENCE

Vincent Divol *

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 some parameters quantifying the regularity of M (such as its reach), 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 $(\hat{M}_t)_{t\geq 0}$, and show that for some choice of t (depending on the regularity parameters), the corresponding estimator is minimax on the class of models of C^2 manifolds introduced in [GPPVW12]. 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. The same selection procedure is then used to design adaptive estimators for tangent spaces and homology groups of the manifold M.

1 Introduction

Manifold inference deals with the estimation of geometric quantites in a random setting. Given $\mathbb{X}_n = \{X_1, \dots, X_n\}$ a set of i.i.d. observations from some law P on $\mathbb{R}^D$ supported on (or concentrated around) a manifold M, one wants to produce an estimator $\hat{\theta}$ which estimates accurately some quantity $\theta(M)$ related to the geometry of M such as its dimension [HA05, LJM09, KRW16], its homology groups [NSW08, BRS+12], its tangent spaces [AL19, CC16], or M itself [GPPVW12, MMS16, AL18, AL19, PS19]. The emphasis has mostly been put on designing estimators attaining minimax rates on a variety of models, which take into account different regularities of the manifold and noise models. Those estimators rely on the knowledge of quantities related either to the geometry of the manifold, such as its dimension or its reach, or to the underlying distribution, such as bounds on its density. Apart from very specific cases, one will not have access to those quantities in practice. One possibility to overcome this issue is to estimate in a preprocessing step those parameters. This may however become the main bottleneck in the estimating process, as regularity parameters are typically harder to estimate than the manifold itself (see for instance [AKC+19] for minimax rates for the estimation of the reach of a manifold).

Another approach, to which this paper is dedicated, consists in designing *adaptive* estimators of $\theta(M)$. An estimator is called adaptive if it attains optimal rates of convergence on a

* *Inria Saclay and Université Paris-Sud*, firstname.lastname@inria.fr

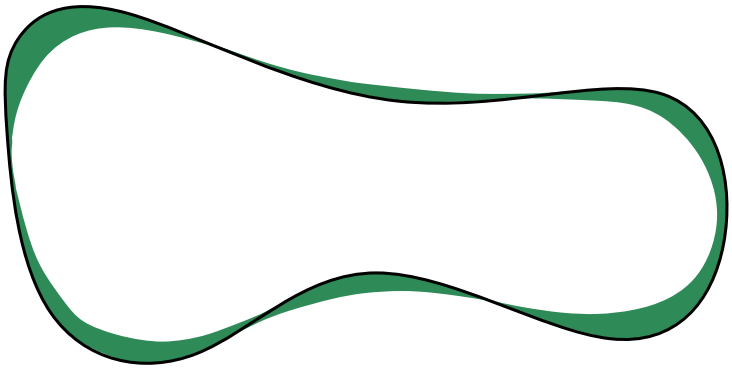


Figure 1 – The t-convex hull $\mathrm{Conv}_t(A)$ (in green) of a curve A (in black).

large class of models. We introduce a manifold estimator $\hat{M}$ which is minimax (with respect to the Hausdorff distance d_H) simultaneously on all the C^2-models with tubular noise introduced in [GPPVW12] and [AL18] (see Section 2 for a precise definition of the models). Our estimator is built by considering a family of estimators given by the t-convex hull $\mathrm{Conv}_t(\mathbb{X}_n)$ of the set of observations $\mathbb{X}_n$. For a given set A, the t-convex hull $\mathrm{Conv}_t(A)$ is defined by

$$\mathrm{Conv}_t(A) := \bigcup_{\sigma \subset A,\ r(\sigma) \leq t} \mathrm{Conv}(\sigma),$$

where $r(\sigma)$ is the *radius* of a set σ, i.e. the radius of the smallest enclosing ball of σ and $\mathrm{Conv}(\sigma)$ is the convex hull of σ (see Definition 3.1). The t-convex hull is an interpolation between the convex hull $\mathrm{Conv}(A)$ of A ($t = +\infty$) and the set A itself ($t = 0$): it gives a "local convex hull" of A at scale t. See Figure 1 for an example.

The loss $d_H(\mathrm{Conv}_t(\mathbb{X}_n), M)$ of the t-convex hull $\mathrm{Conv}_t(\mathbb{X}_n)$ can be efficiently controlled for t larger than some threshold $t^*(\mathbb{X}_n)$ (see Definition 3.4). As the threshold $t^*(\mathbb{X}_n)$ is very close to the sample rate $\varepsilon(\mathbb{X}_n) := d_H(\mathbb{X}_n, M)$ of the point cloud, it is known to be of the order $(\log n/n)^{1/d}$ (see e.g. [RC07, Theorem 2]), and one obtains a minimax estimator on the C^2-models by taking the parameter t of this order (see Theorem 3.7). The exact value of t depends on the unknown parameters of the model (namely the dimension and the reach of the manifold, as well as a lower bound on the density of the distribution), so that it is unclear how the parameter t should be chosen in practice.

The adaptive estimator is build by selecting a parameter $t_\lambda(\mathbb{X}_n)$ (depending on some hyperparameter $\lambda \in (0, 1)$), which is chosen solely based on the observations $\mathbb{X}_n$. More precisely, we consider the convexity defect function of a set A, originally introduced in [ALS13], and defined by

$$h(t, A) = d_H(\mathrm{Conv}_t(A), A) \in [0, t]. \tag{1.1}$$

As its name indicates, the convexity defect function measures how far a set is from being convex at a given scale. For instance, the convexity defect function of a convex set is null, whereas for

a manifold M with positive reach $\tau(M)$, $h(t, M) \leq t^2/\tau(M)$ for $t < \tau(M)$, so that a manifold M is "locally almost convex" (see Proposition 4.2). We show that the convexity defect function of $\mathbb{X}_n$ exhibits a sharp change of behavior around the threshold $t^*(\mathbb{X}_n)$. Namely, for values t which are smaller than a fraction of $t^*(\mathbb{X}_n)$, the convexity defect function $h(t, \mathbb{X}_n)$ has a linear behavior, with a slope approximately equal to 1 (see Proposition 4.3), whereas for $t \geq t^*(\mathbb{X}_n)$, the convexity defect function exhibits the same quadratic behavior than the convexity defect of a manifold (see Proposition 4.4). In particular, its slope is much smaller than 1 as long as $t \geq t^*(\mathbb{X}_n)$ is significantly smaller than the reach $\tau(M)$. This change of behavior at the value $t^*(\mathbb{X}_n)$ suggests to select the parameter

$$t_\lambda(\mathbb{X}_n) := \sup\{t < t_{\max},\ h(t, \mathbb{X}_n) > \lambda t\},$$

where $\lambda \in (0, 1)$ and $t_{\max}$ is a parameter which has to be smaller than the reach $\tau(M)$ of the manifold (see Definition 4.5). We show (see Proposition 4.6) that with high probability, in the case where the sample $\mathbb{X}_n$ is exactly on the manifold M, we have

$$t^*(\mathbb{X}_n) \leq t_\lambda(\mathbb{X}_n) \leq \frac{2t^*(\mathbb{X}_n)}{\lambda}\left(1 + \frac{t^*(\mathbb{X}_n)}{\tau(M)}\right). \tag{1.2}$$

In particular, we are able to control the loss of $\hat{M} := \mathrm{Conv}_{t_\lambda(\mathbb{X}_n)}(\mathbb{X}_n)$ with high probability. By choosing $t_{\max}$ as a slowly decreasing function of n (for instance, $t_{\max} = (\log n)^{-1}$), we obtain an adaptive estimator on the whole collection of C^2-models as defined in Section 2 (see Corollary 4.7 and Remark 4.8 afterwards).

The estimator $\hat{M}$ is to our knowledge the first minimax adaptive, completely data-driven, manifold estimator. Our procedure allows us to actually estimate (up to a multiplicative constant) the sample rate $\varepsilon(\mathbb{X}_n)$. The parameter $t_\lambda(\mathbb{X}_n)$ can therefore be used as an hyperparameter in different settings. To illustrate this general idea, we show how to create an adaptive estimator of the homology groups (see Corollary 4.9) and of the tangent spaces (see Corollary 4.10) of a manifold.

Related work

"Localized" versions of convex hulls such as the t-convex hulls have already been introduced in the support estimation litterature. For instance, slightly modified versions of the t-convex hull have been used as estimators in [AB16] under the assumption that the support has a smooth boundary and in [RC07] 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 [GPPVW12] and [AL18], in which manifold estimators were also proposed. If the estimator in [GPPVW12] is of purely theoretical interest, the estimator proposed by Aamari and Levrard in [AL18], based

on the Tangential Delaunay complex, is computable in polynomal time in the number of inputs and linear in the ambient dimension D. 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 $t_\lambda(\mathbb{X}_n)$ is completely data-driven, while keeping the minimax property. In Section 5, we propose to select a parameter $\tilde{t}_\lambda$ which in practice shares similar properties than $\hat{t}_\lambda$, while being efficiently computable. However, unlike in the case of the Tangential Delaunay complex, we have no guarantees on the homotopy type of the corresponding estimator.

A powerful method to select estimators is given by Lepski's method [Lep92, Bir01] (and its further refinement known as Goldenshluger-Lepski's method, see e.g. [GL13]). In its simplest form, this method applies to a hierarchized family of estimators $(\hat{\theta}_t)_{t\geq 0}$ of some $\theta \in \mathbb{R}$: typically, we assume that the bias of the estimators is a nondecreasing function of t whereas their variance is nonincreasing. The Lepski method consists in comparing each estimator $\hat{\theta}_t$ to the less biased estimators $\hat{\theta}_{t'}$ for $t' \leq t$ and by choosing the smallest t for which the estimator $\hat{\theta}_t$ is close enough to its less biased counterparts (with respect to t). Our method is based on a similar idea, with the important modification that instead of comparing $\hat{\theta}_t$ to all the estimators $\hat{\theta}_{t'}$ for $t' < t$, we show that it is enough to compare each estimator to some degenerate estimator (here corresponding to $\mathbb{X}_n = \mathrm{Conv}_0(\mathbb{X}_n)$) to select a parameter which leads to an adaptive estimator. In that sense, our method largely stems from the Penalized Comparison to Overfitting method introduced in [LMR17] in the framework of kernel density estimation.

Outline of the paper

Notations and preliminary results on manifold estimation are detailed in Section 2. In Section 3, we define the t-convex hull of a set, and show that the estimator $\mathrm{Conv}_t(\mathbb{X}_n)$ is minimax for some choice of t. In Section 4, we introduce the convexity defect function of a set, originally defined in [ALS13], and study in details the behavior of the convexity defect of the observation set $\mathbb{X}_n$. This study is then used to select a parameter $t_\lambda(\mathbb{X}_n)$, depending on two hyperparameters λ and $t_{\max}$, and we show the adaptivity of the estimator $\mathrm{Conv}_{t_\lambda(\mathbb{X}_n)}(\mathbb{X}_n)$. We also discuss how the scale parameter $t_\lambda(\mathbb{X}_n)$ can be used as a scale parameter in the settings of homology and tangent spaces estimation, leading to adaptive procedures in those two frameworks as well. We present some numerical illustrations of our procedure on synthetic datasets in Section 5. A discussion is given in Section 6. Proofs of the main results are found in the Appendix.

2 Preliminaries

Throughout the paper, we fix a probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$ and denote by $\mathbb{E}$ the integration with respect to $\mathbb{P}$. All the random variables X_i, Y_i, Z_i appearing in the following have for domain

this same probabilistic space.

On the use of constants

Except if explicitly stated otherwise, symbols $c_0, c_1, C_0, C_1, \ldots$ will denote absolute constants in the following. If a constant depends on additional parameters $\alpha, \beta, \ldots$, it will be denoted by $C_{\alpha,\beta,\ldots}$. We also write $a \lesssim_{\alpha,\beta,\ldots} b$ for $a \leq C_{\alpha,\beta,\ldots} b$ and $a \asymp_{\alpha,\beta,\ldots} b$ if $a \lesssim_{\alpha,\beta,\ldots} b$ and $b \lesssim_{\alpha,\beta,\ldots} a$.

Notations

Let $\mathcal{C}^2_d$ be the set of $\mathcal{C}^2$ compact connected d-dimensional submanifolds of $\mathbb{R}^D$ without boundary and let $M \in \mathcal{C}^2_d$.

- $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^D$ and $\langle \cdot, \cdot \rangle$ the dot product.
- If $A \subset \mathbb{R}^D$ and $x \in \mathbb{R}^D$, then $d(x, A) := \inf_{y \in A} \|x - y\|$.
- If $A \subset \mathbb{R}^D$, $\operatorname{diam}(A) := \sup_{x,y \in A} \|x - y\|$ is the diameter of A.
- Given $x \in \mathbb{R}^D$ and $r \geq 0$, $\mathcal{B}(x, r)$ is the closed ball of radius r centered at x and, for $A \subset \mathbb{R}^D$, we write $\mathcal{B}_A(x, r)$ for $\mathcal{B}(x, r) \cap A$.
- For $p \in M$, T_pM is the tangent space of M at p. It is identified with a d-dimensional subspace of $\mathbb{R}^D$.
- The asymmetric Hausdorff distance between sets $A, B \subset \mathbb{R}^D$ is defined as

$$d_H(A|B) := \sup_{x \in A} d(x, B). \tag{2.1}$$

  The Hausdorff distance between A and B is then defined as

$$d_H(A, B) = \max\{d_H(A|B), d_H(B|A)\}. \tag{2.2}$$

- For $A \subset M$, we denote by $\varepsilon(A) := d_H(A, M)$ the sample rate of A.

The Hausdorff distance can also be expressed as an ∞-norm between distance functions: for A, B sets of $\mathbb{R}^D$,

$$\sup_{x \in \mathbb{R}^D} |d(x, A) - d(x, B)| = d_H(A, B). \tag{2.3}$$

This directly implies that for any sets $A, B, C \subset \mathbb{R}^D$, one has

$$d_H(A|C) \leq d_H(A|B) + d_H(B, C), \tag{2.4}$$

a fact we will use in the following.

Reach of a manifold

The regularity of a submanifold M 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 an 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 [Fed59], where it is also proven that a $\mathcal{C}^2$ compact submanifold without boundary has a positive reach $\tau(M) > 0$ (see [Fed59, p. 432]). For $\tau_{\min} > 0$, we denote by $\mathcal{C}^2_{d,\tau_{\min}}$ the set of manifolds $M \in \mathcal{C}^2_d$ with reach larger than $\tau_{\min}$.

Minimax rates

Let $\mathcal{P}$ be a set of probability distributions on some measurable space $(\mathcal{X}, \mathcal{G})$ and $\theta : \mathcal{P} \to (E, \rho)$ be a map where (E, ρ) is some metric space. For $n \geq 0$, we denote by $\mathcal{P}^{(n)}$ the set $\{P^{\otimes n},\ P \in \mathcal{P}\}$, where $P^{\otimes n}$ is the product measure of n copies of P. An estimator of θ in $\mathcal{P}^{(n)}$ is any measurable map $\hat{\theta}_n : \mathcal{X}^n \to E$. For $P \in \mathcal{P}$, the quality of the estimator $\hat{\theta}_n$ is measured by its P-risk $\mathbb{E}_P[\rho(\hat{\theta}_n, \theta(P))] := \mathbb{E}[\rho(\hat{\theta}_n(X_1, \ldots, X_n), \theta(P))]$, where $X_1, \ldots, X_n$ is a n-sample of law P. The minimax risk for the estimation of θ on $\mathcal{P}^{(n)}$ is then defined as

$$\mathcal{R}(\theta; \mathcal{P}^{(n)}) := \inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\rho(\hat{\theta}_n, \theta(P))], \tag{2.5}$$

where the infimum is taken on all estimators of θ, i.e. the minimax risk is the best possible risk an estimator can attain uniformly on $\mathcal{P}^{(n)}$. An estimator $\hat{\theta}_n$ realizing this infimum (up to a constant) is called *minimax*.

Rates of convergence of minimax risks as $n \to +\infty$ have been studied in the framework of manifold estimation. Namely, we consider the following models:

Definition 2.1 (Noise-free model)**.** *Let d be an integer smaller than D and $\tau_{\min}$, $f_{\min}$, $f_{\max}$ be positive constants. The set $\mathcal{P}_{d,\tau_{\min},f_{\min},f_{\max}}$ is the set of all distributions having for support a manifold $M \in \mathcal{C}^2_{d,\tau_{\min}}$, which are absolutely continuous with respect to the volume measure on M, and such that their densities with respect to the volume measure are bounded from below by $f_{\min}$ and from above by $f_{\max}$.*

Definition 2.2 (Tubular noise model)**.** *Let d be an integer smaller than D and $\tau_{\min}$, $f_{\min}$, $f_{\max}$, γ be positive constants. We say that $P \in \mathcal{P}_{d,\tau_{\min},f_{\min},f_{\max},\gamma}$ if a random variable X distributed according to P can be written as $X = Y + Z$, where Y and Z are two independent random variables, with the law of Y which is in $\mathcal{P}_{d,\tau_{\min},f_{\min},f_{\max}}$ and with $\|Z\| \leq \gamma$.*

For $P \in \mathcal{P}_{d,\tau_{\min},f_{\min},f_{\max},\gamma}$, let M be the underlying manifold of the distribution P. Then M belongs to the space of all compact subsets of $\mathbb{R}^D$, which is a metric space when endowed with

the Hausdorff distance d_H. Minimax rates for the estimation of the manifold M with respect to the Hausdorff distance in the model $\mathcal{P}_{d,\tau_{\min},f_{\min},f_{\max},\gamma}$ have been studied in [AL18], following the works of [GPPVW12, KZ15].

Theorem 2.3 (Theorem 1 in [KZ15] and Theorem 2.9 in [AL18])**.** *Let $\tau_{\min}, f_{\min}, f_{\max}, \eta$ be positive constants and d be a positive integer smaller than D. For n large enough, if $\gamma_n \leq \eta(\log n/n)^{2/d}$, then*

$$\mathcal{R}(M; \mathcal{P}^{(n)}_{d,\tau_{\min},f_{\min},f_{\max},\gamma_n}) \asymp_{d,\tau_{\min},f_{\min},f_{\max},\eta} \left(\frac{\log n}{n}\right)^{2/d}. \tag{2.6}$$

Probability distributions in Theorem 2.3 contain "almost no noise", as the level of noise γ_n is chosen to be negligible in front of the sample rate $\varepsilon(\mathbb{X}_n)$ (which is of order $(\log n/n)^{1/d}$). Changing the model by adding a small proportion of outliers would not change the minimax rates, as explained in [GPPVW12] or [AL18]. However, the t-convex hull estimators proposed in the next section are very sensible to this addition and some decluttering preprocessing would be needed to obtain better estimators on such models. Note also that the t-convex hull estimators will be minimax on the model

$$\mathcal{P}_{d,\tau_{\min},f_{\min},\gamma_n} := \mathcal{P}_{d,\tau_{\min},f_{\min},+\infty,\gamma_n},$$

for which the minimax rate[1] is also equal to $(\log n/n)^{2/d}$.

3 Minimax manifold estimation with t-convex hulls

Let $\sigma \subset \mathbb{R}^D$. There exists an unique closed ball with minimal radius which contains σ (see [ALS13, Lemma 15]). This ball is called the *minimal enclosing ball* of σ and its radius, called the radius of σ, is denoted by $r(\sigma)$ in the following.

Definition 3.1. *Let $A \subset \mathbb{R}^D$ and $t \geq 0$. The t-convex hull of A is defined as*

$$\mathrm{Conv}_t(A) := \bigcup_{\sigma \subset A,\ r(\sigma) \leq t} \mathrm{Conv}(\sigma). \tag{3.1}$$

In this section, we derive rates of convergence for the estimators $\mathrm{Conv}_t(\mathbb{X}_n)$, where $\mathbb{X}_n$ is a n-sample from law $P \in \mathcal{P}_{d,\tau_{\min},f_{\min},\gamma_n}$.

Remark 3.2. The application taking its values in the space of compact subsets of $\mathbb{R}^D$ endowed with its Borel σ-field and defined by:

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

[1] The minimax rate is lower bounded by $(\log n/n)^{2/d}$, as the model is larger than $\mathcal{P}_{d,\tau_{\min},f_{\min},1,\gamma_n}$ for which the rate is known to be $(\log n/n)^{2/d}$. The study of the estimator in Section 3 will show the upper bound.

is measurable. Indeed, it can be written as $\bigcup_{I\subset\{1,\dots,n\}} \operatorname{Conv}(\{x_i\}_{i\in I}) \cap f_I(x_1,\dots,x_n)$ where $f_I(x_1,\dots,x_n) = \emptyset$ if $r(\{x_i\}_{i\in I}) > t$ and is equal to $\mathbb{R}^D$ otherwise. As the operations $\cup$, $\cap$, Conv are measurable (see [Aam17, Proposition III.7]) and the function r is continuous (see [ALS13, Lemma 16]), the measurability follows.

To obtain rates of convergence, we bound the Hausdorff distance $d_H(\operatorname{Conv}_t(A), M)$ for a general subset $A \subset M$. First, [ALS13, Lemma 12] gives a bound on the asymmetric Hausdorff distance between the convex hull of a subset of M and the manifold M.

Lemma 3.3. *Let $\sigma \subset M$ with $r(\sigma) < \tau(M)$ and let $y \in \operatorname{Conv}(\sigma)$. Then,*

$$d(y, M) \leq \frac{r(\sigma)^2}{\tau(M)}. \tag{3.2}$$

Proof. Lemma 12 in [ALS13] states that if $\sigma \subset M$ satisfies $r(\sigma) < \tau(M)$ and $y \in \operatorname{Conv}(\sigma)$, then,

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

As $\sqrt{u} \geq u$ for $u \in [0, 1]$, one obtains the conclusion. □

This lemma directly implies that $d_H(\operatorname{Conv}_t(A)|M) \leq t^2/\tau(M)$ if $t < \tau(M)$, so that the set $\operatorname{Conv}_t(A)$ is included in the t-neighborhood of M. Therefore, the projection π_M is well defined on the t-convex hull of A for such a t. We introduce a scale parameter $t^*(A)$, which has to be thought as the "best" scale parameter t for approximating M with $\operatorname{Conv}_t(A)$.

Definition 3.4. *For $A \subset M$, let*

$$t^*(A) := \inf\{t < \tau(M),\ \pi_M(\operatorname{Conv}_t(A)) = M\} \in [0, \tau(M)) \cup \{+\infty\}. \tag{3.3}$$

See Figure 2 for an illustration. Assume that $t^*(A) < +\infty$. Then, for $t^*(A) < t < \tau(M)$, and for any point $p \in M$, there exists $y \in \operatorname{Conv}_t(A)$ with $\pi_M(y) = p$. Therefore,

$$d(p, \operatorname{Conv}_t(A)) \leq \|y - p\| = d(y, M) \leq d_H(\operatorname{Conv}_t(A)|M).$$

By taking the supremum over $p \in M$, we obtain that for any $t^*(A) < t < \tau(M)$.

$$\begin{aligned} d_H(\operatorname{Conv}_t(A), M) &= \max\{d_H(\operatorname{Conv}_t(A)|M), d_H(M|\operatorname{Conv}_t(A))\\ &= d_H(\operatorname{Conv}_t(A)|M) \leq \frac{t^2}{\tau(M)}. \end{aligned} \tag{3.4}$$

The minimax rate is now obtained thanks to two observations: (i) $t^*(A)$ is close to the sample rate $\varepsilon(A)$ and (ii) the sample rate of a random sample can be very well controlled.

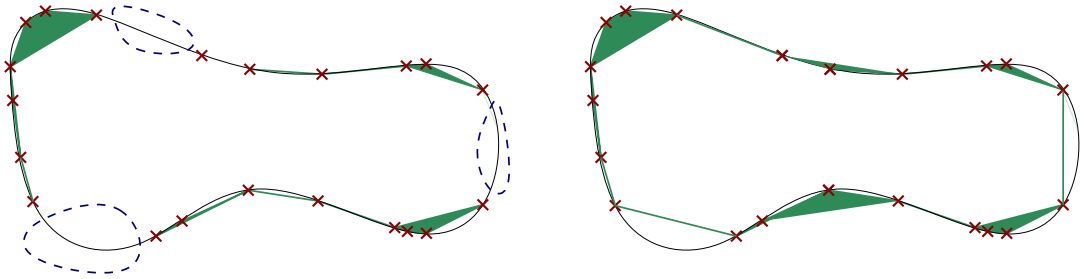


Figure 2 – The t-convex hull of the finite set A (red crosses) is displayed (in green) for two values of t. The black curve represents the (one dimensional) manifold M. On the first display, the value of t is smaller than $t^*(A)$, as there are regions of the manifold (circled in blue) which are not 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.

Proposition 3.5. *Let $A \subset M$ be a finite set. If $\varepsilon(A) \leq C_0 \tau(M)$, then*

$$\varepsilon(A)\left(1 - C_1 \frac{\varepsilon(A)}{\tau(M)}\right) \leq t^*(A) \leq \varepsilon(A)\left(1 + C_2 \frac{\varepsilon(A)}{\tau(M)}\right). \tag{3.5}$$

In particular, $t^*(A)$ is finite.

The proof of Proposition 3.5 is found in Appendix B.1.

Proposition 3.6 (Lemma III.23 in [Aam17])**.** *Let $P \in \mathcal{P}_{d,\tau_{\min},f_{\min}}$ and $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ a n-sample of law P. If $r \leq \tau_{\min}/2$, then*

$$\mathbb{P}(\varepsilon(\mathbb{X}_n) > r) \leq \frac{c_d}{f_{\min} r^d} \exp(-n C_d f_{\min} r^d). \tag{3.6}$$

By gathering those different observations and by using stability properties of t-convex hulls with respect to noise, we show that t-convex hulls are minimax estimators on C^2-models.

Theorem 3.7. *Let d be an integer smaller than D and $\tau_{\min}, f_{\min}, \eta > 0$. Then, for n large enough, and $\gamma_n \leq \eta\,(\log n/n)^{2/d}$, by letting $t_n = C_{d,\tau_{\min},f_{\min}}\,(\log n/n)^{1/d}$, we have*

$$\sup_{P \in \mathcal{P}_{d,\tau_{\min},f_{\min},\gamma_n}} \mathbb{E}_P d_H(\,\mathrm{Conv}_{t_n}(\mathbb{X}_n), M) \lesssim_{d,\tau_{\min},f_{\min},\eta} \left(\frac{\log n}{n}\right)^{2/d}, \tag{3.7}$$

i.e. $\mathrm{Conv}_{t_n}(\mathbb{X}_n)$ is a minimax estimator of M on $\mathcal{P}^{(n)}_{d,\tau_{\min},f_{\min},\gamma_n}$.

A proof of Theorem 3.7 is found in Appendix B.2.

4 Selection procedure for the t-convex hulls

Assuming that we have observed a n-sample $\mathbb{X}_n$ having a distribution $P \in \mathcal{P}_{d,\tau_{\min},f_{\min},\gamma_n}$, 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 $d, \tau_{\min}, f_{\min}, \gamma_n$, whereas those quantities will likely not be accessible in practice. A powerful idea to overcome this issue is to design a selection procedure for the family of estimators $(\operatorname{Conv}_t(\mathbb{X}_n))_{t\geq 0}$. Assume first for the sake of simplicty that *the noise level γ_n is null*. As the loss of the estimator $\operatorname{Conv}_t(\mathbb{X}_n)$ is controlled efficiently for $t \geq t^*(\mathbb{X}_n)$ (see (3.4)), a good idea is to select the parameter t larger than $t^*(\mathbb{X}_n)$. We however do not have access to this quantity based on the observations $\mathbb{X}_n$, as the manifold M is unknown. To select a scale close to $t^*(\mathbb{X}_n)$, we monitor how the estimators $\operatorname{Conv}_t(\mathbb{X}_n)$ deviate from $\mathbb{X}_n$ as t increases. Namely, we use the convexity defect function introduced in [ALS13].

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

$$h(t, A) := d_H(\operatorname{Conv}_t(A), A). \tag{4.1}$$

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 $A \subset \mathbb{R}^D$ be a closed set and $t \geq 0$.*

1. *We have $0 \leq h(t, A) \leq t$.*
2. *A is convex if and only if $h(\cdot, A) \equiv 0$.*
3. *If M is a manifold of reach $\tau(M)$ and $t < \tau(M)$, then*

$$h(t, M) \leq t^2/\tau(M). \tag{4.2}$$

Proof. Point 1 is stated in [ALS13, Section 3.1], Point 2 is clear and Point 3 is a consequence of Lemma 3.3. □

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 convex).

The convexity defect function $h(\cdot, \mathbb{X}_n)$ of the set of observations $\mathbb{X}_n$ has two very different behaviors according to the values of t, as summed up by the two following propositions.

Proposition 4.3 (Short-scale behavior)**.** *Let d be an integer smaller than D, and let $\tau_{\min}$, $f_{\min}$, $f_{\max} > 0$. Let $\mathbb{X}_n$ be a n-sample of law $P \in \mathcal{P}_{d,\tau_{\min},f_{\min},f_{\max}}$. Fix $0 < \lambda < 1$. There exist*

positive constants $t_0, C_0, \dots, C_4$ depending on the parameters of the model and on λ such that the following holds. Let, for $x > 0$, $\phi(x) = C_0 x^2 e^{-C_1 x}$ with constants chosen so that $\phi(x) < 1$. Define $\psi(x) = \phi(x)/\log(1/\phi(x))$. Then, for n large enough and $0 < t \le t_0$, we have

$$h(t, \mathbb{X}_n) \ge \lambda t \text{ with probabilty larger than } 1 - C_2 \exp(-C_3 t^{-d} \psi(C_4 n t^d)). \tag{4.3}$$

The proof of Proposition 4.3 is found in Appendix C.1.

Proposition 4.4 (Long-scale behavior)**.** *Let $A \subset M$. For $t^*(A) < t < \tau(M)$,*

$$h(t, A) \le \frac{t^2}{\tau(M)} + t^*(A) \left(1 + \frac{t^*(A)}{\tau(M)}\right). \tag{4.4}$$

Proof. By using that $h(t, A) \le t$ and (3.4), for any $t^*(A) < s < t$,

$$\begin{aligned} h(t, A) &= d_H(\operatorname{Conv}_t(A), A) \\ &\le d_H(\operatorname{Conv}_t(A), M) + d_H(M, \operatorname{Conv}_s(A)) + d_H(\operatorname{Conv}_s(A), A) \\ &\le \frac{t^2}{\tau(M)} + \frac{s^2}{\tau(M)} + s. \end{aligned}$$

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

Let us shortly explain the content of the two previous propositions. The probability appearing in (4.3) will be close to 1 as long as t is smaller than a fraction of $(\log n/n)^{1/d}$ and larger than $(1/n)^{(2-\delta)/d}$ for any $0 < \delta < 1$. Therefore, with high probability, the convexity defect function $h(t, \mathbb{X}_n)$ is very close to t for $(1/n)^{(2-\delta)/d} \lesssim t \lesssim (\log n/n)^{1/d}$. On the contrary, standard techniques show that if $t \lesssim (1/n)^{2/d}$, then $h(t, \mathbb{X}_n)$ is null with probability larger than, say, $1/2$, indicating that the lower bound in the previous range is close of being optimal. The arguments to prove Proposition 4.3 are of a purely probabilistic nature and do not rely on the geometry of the support of P. On the contrary, the long-scale behavior described in Proposition 4.4 relies only on the geometry of M and is completely deterministic, in the sense that it holds for any set $A \subset M$ for which $t^*(A) < \tau(M)$. It indicates that when t is larger than the threshold $t^*(A)$ (which is of order $(\log n/n)^{1/d}$ for $A = \mathbb{X}_n$), then the geometry of M becomes the only factor driving the growth of $h(t, A)$, and this growth is the same than the growth of the convexity defect of the manifold M. See also Figure 3.

The previous discussion indicates to choose the smallest t in the quadratic behavior range to select a value larger than (but close to) $t^*(\mathbb{X}_n)$.

Definition 4.5. *Let $A \subset M$, $\lambda > 0$ and $t_{\max} > 0$. We define*

$$t_\lambda(A) := \sup\{t < t_{\max},\ h(t, A) \ge \lambda t\}. \tag{4.5}$$

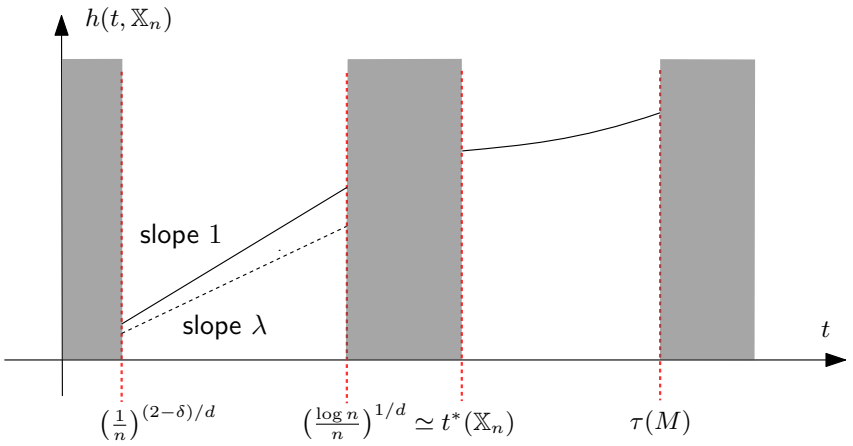


Figure 3 – Summary of the behavior of the convexity defect function of a n-sample $\mathbb{X}_n$ of law $P \in \mathcal{P}_{d,\tau_{\min},f_{\min},f_{\max}}$. Fix $0 < \lambda, \delta < 1$. According to Proposition 4.3, with high probability, the convexity defect function $h(t, \mathbb{X}_n)$ is larger than λt for $(1/n)^{(2-\delta)/d} \lesssim t \lesssim (\log n/n)^{1/d}$. For $t^*(\mathbb{X}_n) < t < \tau(M)$, it exhibits at most quadratic growth by Proposition 4.4 (with $t^*(\mathbb{X}_n)$ of order $(\log n/n)^{1/d}$ as well). The behavior of the convexity defect function for $t \geq \tau(M)$ depends on global properties of the geometry of the support M and cannot be straighforwardly inferred.

Propositions 4.3 and 4.4 prove that for any $\lambda < 1$, $t_\lambda(\mathbb{X}_n)$ is with high probability of order $(\log n/n)^{1/d}$, that is of the same order than $t^*(\mathbb{X}_n)$. However, selecting a parameter of the order of $t^*(\mathbb{X}_n)$ is not enough to obtain a tight bound on the loss, as such a control only holds for $t > t^*(\mathbb{X}_n)$ (at least in the noise-free model, see (3.4)). We are able to obtain a more precise inequality for general subsets B close to M, as summed up by the next proposition.

Theorem 4.6. *Let $0 < \lambda < 1$, $\gamma \geq 0$ and $M \in \mathcal{C}^2_d$. Let $A \subset M$ be a finite set with $\varepsilon(A) \leq C_1\tau(M)$ and $B \subset \mathbb{R}^D$ with $d_H(A, B) \leq \gamma$. Assume that*

1. *$t^*(A) + \gamma < t_{\max} < \tau(M)\lambda/2 - \gamma$,*
2. *$t^*(A) < C_2(1-\lambda)\tau(M)$ and $t^*(A) \leq C_3\lambda^2\tau(M)$,*
3. *$\gamma \leq C_4(1-\lambda)t^*(A)$.*

Then,

$$t^*(A) + \gamma \leq t_\lambda(B) \leq \frac{2t^*(A)}{\lambda}\left(1 + \frac{t^*(A)}{\tau(M)}\right) + \frac{6\gamma}{\lambda}. \tag{4.6}$$

The proof of Theorem 4.6 is found in Appendix C.2. As a corollary of this result, we obtain the adaptivity of the the t-convex hull estimators of parameter $t_\lambda(\mathbb{X}_n)$.

Corollary 4.7. *Let $0 < \lambda < 1$ and $t_{\max} > 0$. Let d be an integer smaller than D and $f_{\min}, \eta > 0$, $\tau_{\min} > 2t_{\max}/\lambda$. Then, for n large enough, and $\gamma_n \leq \eta\,(\log n/n)^{2/d}$, we have*

$$\sup_{P \in \mathcal{P}_{d,\tau_{\min},f_{\min},\gamma_n}} \mathbb{E}_P d_H(\,\mathrm{Conv}_{t_\lambda(\mathbb{X}_n)}(\mathbb{X}_n), M) \lesssim_{d,\tau_{\min},f_{\min},\eta,\lambda,t_{\max}} \left(\frac{\log n}{n}\right)^{2/d}, \tag{4.7}$$

i.e. the estimator $\mathrm{Conv}_{t_\lambda(\mathbb{X}_n)}(\mathbb{X}_n)$ is adaptive minimax on all the models $\mathcal{P}^{(n)}_{\tau_{\min},f_{\min},d,\gamma_n}$ for $d > 0$, $f_{\min} > 0$, $\tau_{\min} > 2t_{\max}/\lambda$ and $\gamma_n \leq \eta\,(\log n/n)^{2/d}$ for some $\eta > 0$.

A proof of Corollary 4.7 is found in Appendix C.3.

Remark 4.8. From an asymptotic perspective, one may simply set $t_{\max} = (\log n)^{-1}$ to obtain an estimator which is simultaneously minimax on all models $\mathcal{P}_{d,\tau_{\min},f_{\min},\gamma_n}$ with $\tau_{\min} > 0$. Taking such an approach may obscure the fact that, should n be not large enough (i.e. if $t^*(\mathbb{X}_n)$ is larger than some fraction of the reach), then the selection procedure is doomed to fail, as the long-scale behavior corresponding to the range $[t^*(\mathbb{X}_n), \tau(M)]$ is too small to be captured by the selection procedure (or even is non-existent).

Another possible criterion to ensure the quality of an estimator $\hat{M}$ of a manifold M is to ensure that $\hat{M}$ and M are homotopy equivalent. Although we have no guarantees on the topology of the estimator $\mathrm{Conv}_{t_\lambda(\mathbb{X}_n)}(\mathbb{X}_n)$, our selection procedure also permits to build a simplicial complex homotopy equivalent to M. We write $M \simeq N$ to indicate that the two topological spaces M and N are homotopy equivalent. For $A \subset \mathbb{R}^D$, the Čech simplicial complex of parameter t on A is defined as

$$\mathcal{C}_t(A) := \{\sigma \subset A,\ r(\sigma) \leq t\}. \tag{4.8}$$

We will consider that $\mathcal{C}_t(A)$ is a topological space by identifying it with its geometric realization.

Corollary 4.9. *Let $0 < \lambda < 1$ and $t_{\max} > 0$. Let d be an integer smaller than D and $f_{\min}, \eta > 0$, $\tau_{\min} > 2t_{\max}/\lambda$. Then, for n large enough, and $\gamma_n \leq \eta\,(\log n/n)^{2/d}$, we have*

$$\sup_{P \in \mathcal{P}_{d,\tau_{\min},f_{\min},\gamma_n}} \mathbb{P}(M \not\simeq \mathcal{C}_{5t_\lambda(\mathbb{X}_n)}(\mathbb{X}_n)) \lesssim_{d,\tau_{\min},f_{\min},\eta,\lambda,t_{\max}} \exp(-C_{d,\tau_{\min},f_{\min},\eta,\lambda,t_{\max}} n). \tag{4.9}$$

This rate matches the exponential minimax rate obtained in [BRS+12] for estimating homology groups, i.e. the parameter $t_\lambda(\mathbb{X}_n)$ also allows to create adaptive minimax homology estimators.

As a last example, we show that the parameter $t_\lambda(\mathbb{X}_n)$ can also be used to estimate tangent spaces in an adaptive way. Let $p \in M$ and $A \subset M$ be a finite set. We denote by $T_p(A, t)$ to be the d-dimensional vector space U which minimizes $d_H(A \cap \mathcal{B}(p, t), p + U)$. This estimator was originally studied in [BSW09]. The angle between subspaces is denoted by $\angle$ (see Appendix A).

Corollary 4.10. *Let $0 < \lambda < 1$ and $t_{\max} > 0$. Let d be an integer smaller than D and $f_{\min} > 0$, $\tau_{\min} > 2t_{\max}/\lambda$. Then, for n large enough, we have*

$$\sup_{P \in \mathcal{P}_{d,\tau_{\min},f_{\min}}} \mathbb{E}\angle(T_pM, T_p(\mathbb{X}_n, 11t_\lambda(\mathbb{X}_n))) \lesssim_{d,\tau_{\min},f_{\min},\lambda,t_{\max}} \left(\frac{\log n}{n}\right)^{1/d}. \tag{4.10}$$

This rate is the minimax rate (up to logarithmic factors) according to [AL19, Theorem 3]. Proofs of Corollary 4.9 and 4.10 are found in Appendix C.3.

A remark on the choice of the scale function in the definition of the t-convex hulls

The study of Sections 3 and 4 was conducted with the t-convex hulls defined with the radius $r(\sigma)$ of a set σ. More generally, one can consider a function $\rho : \mathcal{F}(\mathbb{R}^D) \to \mathbb{R}$, where $\mathcal{F}(\mathbb{R}^D)$ is the powerset of $\mathbb{R}^D$, and define

$$\mathrm{Conv}_t^\rho(A) = \bigcup_{\sigma \subset A,\ \rho(\sigma) \leq t} \mathrm{Conv}(\sigma). \tag{4.11}$$

Assume that there exists two constants a, b such that $ar(\sigma) \leq \rho(\sigma) \leq br(\sigma)$. Then,

$$\mathrm{Conv}_{t/b}(A) \subset \mathrm{Conv}_t^\rho(A) \subset \mathrm{Conv}_{t/a}(A). \tag{4.12}$$

This interleaving between the two filtrations of sets directly implies that (up to straightforward modifications) any result on the asymptotic behavior of $\mathrm{Conv}_t(\mathbb{X}_n)$ or $t_\lambda(\mathbb{X}_n)$ can be translated on its ρ counterparts. As an example, one may choose $\rho = \mathrm{diam}$, as Jung's theorem implies that the bi-Lipschitz condition holds with constants $a = \sqrt{2},\ b = 2$.

5 Numerical considerations

The selection procedure described in Section 4 amounts to compute the convexity defect function of the set $\mathbb{X}_n$, which itself amounts to compute the Hausdorff distance between a family of simplexes and a point cloud in dimension D. The time complexity of this problem is (naively) of order n^D where n is the number of points, and is therefore not tractable in high dimension. To overcome this problem, we propose to modify slightly the definition of $t_\lambda(\mathbb{X}_n)$ by defining

$$\mathrm{Graph}_t(A) := \bigcup_{x_1,x_2 \in A,\ \|x_1 - x_2\| \leq 2t} \mathrm{Conv}(\{x_1, x_2\}), \tag{5.1}$$

$$\tilde{h}(t, A) := d_H(\mathrm{Graph}_t(A), A) \text{ and} \tag{5.2}$$

$$\tilde{t}_\lambda(A) := \sup\{t < t_{\max},\ \tilde{h}(t, A) \geq \lambda t\}. \tag{5.3}$$

The function $\tilde{h}(t, A)$ can be computed efficiently. Indeed, for each edge $e = \{x_1, x_2\} \subset A$, the distance $d_H(\mathrm{Conv}(e)|A)$ can be computed by considering the Delaunay triangulation of the set

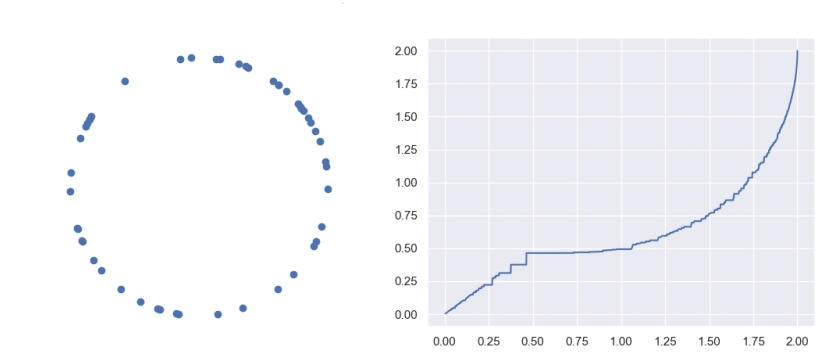
$\{(\pi_L(x), d(x,L)),\ x \in A\} \subset \mathbb{R}^2$, where L is the line passing through the two points of e. It can therefore be computed in a $O(nD + n\log n)$ time, where n is the cardinality of A. This time can be further reduced by considering only points which are in a neighborhood of the edge e instead of the n points of the data set. As there are n^2 edges in the dataset, a crude upperbound on the complexity of the computation of $\tilde{t}_\lambda(A)$ is $O(n^2(nD + n\log n))$. Note that we have no theoretical guarantees on the parameter $\tilde{t}_\lambda(A)$. However, numeric experiments show that $\tilde{h}(t, \mathbb{X}_n)$ exhibits a behavior similar to the one of $h(t, \mathbb{X}_n)$.

In Figure 4, we compute the convexity defect function $\tilde{h}(t, \mathbb{X}_n)$ of three synthetic datasets: (a) $n = 50$ points uniformly sampled on a circle, (b) $n = 500$ points uniformly sampled on a torus, and (c) $n = 5000$ points sampled on a swissroll. On each convexity defect function, the behavior described in Section 4 is observed: first a linear growth up to a certain value, then a quadratic growth, and eventually a sharp change of behavior at the reach (equal to 2 in the first two illustrations, and slightly larger than 6 on the swiss roll dataset). Moreover, we also computed the homology groups of the Čech complex, with a parameter selected in the quadratic regime of the convexity defect function (respectively $t = 0.5$, $t = 1.3$ and $t = 3.5$ on datasets (a), (b) and (c)). Each time, the homology groups of the Čech complex coincide with the homology groups of the underlying manifold. This is expected, as the proposed parameters t are likely to be very close to $t^*(\mathbb{X}_n)$, and larger than the sample rate $\varepsilon(\mathbb{X}_n)$.

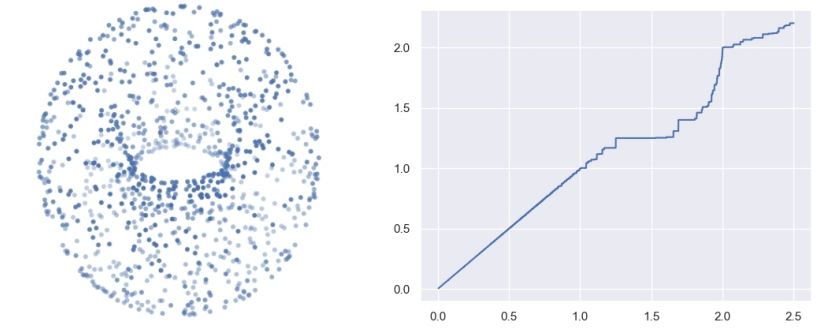
6 Discussion and further works

In this article, we introduced a particularly simple manifold estimator, based on an 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. Surprisingly enough, the selection procedure allows to find a parameter $t_\lambda(\mathbb{X}_n)$ which is with high probability between, say, $\varepsilon(\mathbb{X}_n)/2$ and $2\varepsilon(\mathbb{X}_n)$ (at least for 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 adaptive tangent spaces and homology estimators can be created thanks to $\hat{t}_\lambda(\mathbb{X}_n)$. The main limitation to our procedure is its non-robustness to outliers. Indeed, even in the presence of one outlier, the loss function $t \mapsto d_H(\mathrm{Conv}_t(\mathbb{X}_n), M)$ would be constant, equal to the distance between the outlier and the manifold M: with respect to the Hausdorff distance, all the estimators $\mathrm{Conv}_t(\mathbb{X}_n)$ are 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 the distance to measure or Wasserstein distances on judicious probability measures built on the t-convex hulls $\mathrm{Conv}_t(\mathbb{X}_n)$ instead of the Hausdorff distance. Other challenges raised by this work include the following:

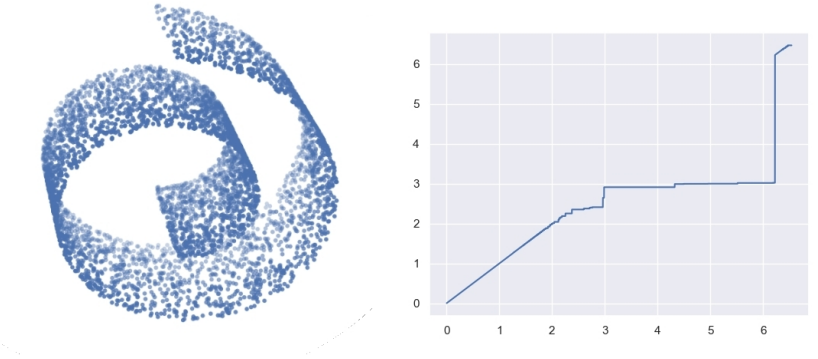
- Besides being close to M for the Hausdorff metric, a desirable property for a manifold



(a)



(b)



(c)

Figure 4 – The convexity defect function of three datasets: $n = 50$ points uniformly sampled on a circle, (b) $n = 500$ points uniformly sampled on a torus, and (c) $n = 5000$ points sampled on a swissroll.

estimator $\hat{M}$ is to be homotopy equivalent to M with high probability. It is an open question whether $\mathrm{Conv}_t(\mathbb{X}_n)$ satisfies this property for some values of t. Note however that it is proven in [AM19, Corollary 5.6] that $\mathrm{Conv}_t(M)$ is homotopy equivalent to M for t smaller than the reach $\tau(M)$.

- Other classes of statistical models, corresponding to manifolds of regularity $k > 2$, were considered in [AL19], and minimax rates of manifold estimation are of order $(\log n/n)^{k/d}$ on those models. A natural challenge raised by the present work is to create an adaptive estimator on all the models of C^k manifolds for $k \geq 2$, i.e. to create an estimator which adapts to the regularity of the manifold as well as to its reach and to other parameters.

Acknowledgements

I am grateful to Fréderic Chazal (Inria Saclay) and Pascal Massart (Université Paris-Sud) for helpful discussions and valuable comments on both mathematical and computational aspects of this work. I would also like to thank Théo Lacombe for his thorough re-reading of the paper. This work was partially supported by the advanced Grant of the European Research Council GUDHI (Geometric Understanding in Higher Dimensions).

References

[Aam17] Eddie Aamari. *Vitesses de convergence en inférence géométrique*. PhD thesis, Paris Saclay, 2017.

[AB16] Catherine Aaron and Olivier Bodart. Local convex hull support and boundary estimation. *Journal of Multivariate Analysis*, 147:82–101, 2016.

[ACLZ17] Ery Arias-Castro, Gilad Lerman, and Teng Zhang. Spectral clustering based on local PCA. *The Journal of Machine Learning Research*, 18(1):253–309, 2017.

[AKC+19] Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the reach of a manifold. *Electronic Journal of Statistics*, 13(1):1359–1399, 2019.

[AL18] Eddie Aamari and Clément Levrard. Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction. *Discrete & Computational Geometry*, 59(4):923–971, 2018.

[AL19] Eddie Aamari and Clément Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47(1):177–204, 2019.

[ALS13] Dominique Attali, André Lieutier, and David Salinas. Vietoris–Rips complexes also provide topologically correct reconstructions of sampled shapes. *Computational Geometry*, 46(4):448–465, 2013.

[AM19] Henry Adams and Joshua Mirth. Metric thickenings of Euclidean submanifolds. *Topology and its Applications*, 254:69–84, 2019.

[Bir01] Lucien Birgé. An alternative point of view on Lepski's method. *Lecture Notes-Monograph Series*, 36:113–133, 2001.

[BRS+12] Sivaraman Balakrishnan, Alesandro Rinaldo, Don Sheehy, Aarti Singh, and Larry Wasserman. Minimax rates for homology inference. In *Artificial Intelligence and Statistics*, pages 64–72, 2012.

[BSW09] Mikhail Belkin, Jian Sun, and Yusu Wang. Constructing Laplace operator from point clouds in $\mathbb{R}^d$. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pages 1031–1040. Society for Industrial and Applied Mathematics, 2009.

[CC16] Siu-Wing Cheng and Man-Kwun Chiu. Tangent estimation from point samples. *Discrete & Computational Geometry*, 56(3):505–557, 2016.

[CCSL09] Frédéric Chazal, David Cohen-Steiner, and André Lieutier. A sampling theory for compact sets in Euclidean space. *Discrete & Computational Geometry*, 41(3):461–479, 2009.

[Fed59] Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.

[Gal] Robert G Gallager. *Information theory and reliable communication*, volume 2. Springer.

[GL13] Alexander Goldenshluger and Oleg Lepski. General procedure for selecting linear estimators. *Theory of Probability and Its Applications*, 57(2):209–226, 2013.

[GPPVW12] Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. *The Annals of Statistics*, 40(2):941–963, 2012.

[HA05] Matthias Hein and Jean-Yves Audibert. Intrinsic dimensionality estimation of submanifolds in $\mathbb{R}^d$. In *Proceedings of the 22nd international conference on Machine learning*, pages 289–296. ACM, 2005.

[KRW16] Jisu Kim, Alessandro Rinaldo, and Larry Wasserman. Minimax rates for estimating the dimension of a manifold. *arXiv preprint arXiv:1605.01011*, 2016.

[KZ15] Arlene KH Kim and Harrison H Zhou. Tight minimax rates for manifold estimation under Hausdorff loss. *Electronic Journal of Statistics*, 9(1):1562–1582, 2015.

[Lep92] Oleg Lepskii. Asymptotically minimax adaptive estimation. I: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1992.

[LJM09] Anna V Little, Yoon-Mo Jung, and Mauro Maggioni. Multiscale estimation of intrinsic dimensionality of data sets. In *2009 AAAI Fall Symposium Series*, 2009.

[LL14] Nathan Linial and Zur Luria. Chernoff's inequality-a very elementary proof. *arXiv preprint arXiv:1403.7739*, 2014.

[LMR17] Claire Lacour, Pascal Massart, and Vincent Rivoirard. Estimator selection: a new method with applications to kernel density estimation. *Sankhya A*, 79(2):298–335, 2017.

[MMS16] Mauro Maggioni, Stanislav Minsker, and Nate Strawn. Multiscale dictionary learning: non-asymptotic bounds and robustness. *The Journal of Machine Learning Research*, 17(1):43–93, 2016.

[NSW08] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008.

[PS19] Nikita Puchkin and Vladimir Spokoiny. Structure-adaptive manifold estimation. *arXiv preprint arXiv:1906.05014*, 2019.

[RC07] Alberto Rodríguez Casal. Set estimation under convexity type assumptions. In *Annales de l'IHP Probabilités et statistiques*, volume 43, pages 763–774, 2007.

A Properties of manifolds with reach constraints

In this section, M is a manifold in $\mathcal{C}^2_d$. We recall that for $p \in M$, T_pM is the tangent space of M at p. The corresponding affine subspace passing through p is denoted by $\tilde{T}_pM$. For $U \subset \mathbb{R}^D$ a vector space, we let π_U be the orthogonal projection on U and $\pi_U^\perp$ be the orthogonal projection on the orthogonal space $U^\perp$. Also, we write π_p for π_{T_pM} and we define $\tilde{\pi}_p : \mathbb{R}^D \to \tilde{T}_pM$ by $\tilde{\pi}_p(x) = \pi_p(x-p) + p$ for $x \in \mathbb{R}^D$, so that $\tilde{\pi}_p(p) = p$.

The angle $\angle(U, V)$ between two subspaces U, V of $\mathbb{R}^D$ is defined as $\|\pi_U - \pi_V\|_{\text{op}}$, the distance for the operator norm between the orthogonal projections on U and V. The principal angle $\theta(U, V)$ is defined by the relation

$$\sin\theta(U,V) := \left\|\pi_V^\perp \circ \pi_U\right\|_{\text{op}}. \tag{A.1}$$

If U and V have the same dimension, then $\sin\theta(U,V) = \angle(U,V)$ (see for instance [Aam17, Section III.4]).

Lemma A.1 (Lemma 3.4 in [BSW09])**.** *Let $p, q \in M$ with $\|p-q\| \leq \tau(M)/2$. Then,*

$$\cos\theta(T_pM, T_qM) \geq 1 - 2\frac{\|p-q\|^2}{\tau(M)^2}.$$

In particular,

$$\angle(T_pM, T_qM) < 2\frac{\|p-q\|}{\tau(M)}.$$

The following characterization of the reach is useful to control how points on manifold deviate from their projections on some tangent space.

Lemma A.2 (Theorem 4.18 in [Fed59])**.** *For $p, q \in M$,*

$$\|\pi_p^{\perp}(q-p)\| \leq \frac{\|q-p\|^2}{2\tau(M)}. \tag{A.2}$$

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

Lemma A.3. *Let $p \in M$.*

1. *Let $x \in \mathbb{R}^D$ with $d(x,M) < \tau(M)$. Then, $\pi_M(x) = p$ if and only if $\tilde{\pi}_p(x) = p$.*
2. *For $r \leq \tau(M)/3$, the application $\tilde{\pi}_p$ is a diffeormorphism from $\mathcal{B}_M(p,r)$ on its image. Moreover, its image $\tilde{\pi}_p(\mathcal{B}_M(p,r))$ contains $\mathcal{B}_{\tilde{T}_pM}(p, 7r/8)$. In particular, if $x \in \mathcal{B}_M(p, \tau(M)/4)$, then*

$$\|\tilde{\pi}_p(x) - p\| \geq \frac{7}{8}\|x-p\|. \tag{A.3}$$

Proof. 1. See Point (12) in [Fed59, Theorem 4.8].

2. We first show that $\tilde{\pi}_p$ is injective on $\mathcal{B}_M(p, \tau(M)/3)$. Assume that $\tilde{\pi}_p(q) = \tilde{\pi}_p(q')$ for some $q \neq q' \in M$. Consider without loss of generality that $\|p-q\| \geq \|p-q'\|$. The goal is to show that $\|p-q\| > \tau(M)/3$. If $\|p-q\| > \tau(M)/2$, the conclusion obviously holds. Lemma

A.1 states that if it is not the case then, $\angle(T_pM, T_qM) < 2\frac{\|p-q\|}{\tau(M)}$. Also, by definition,

$$\begin{aligned}
\angle(T_pM, T_qM) &\geq \frac{\|(\pi_p - \pi_q)(q-q')\|}{\|q-q'\|} \\
&= \frac{\|\pi_q(q-q')\|}{\|q-q'\|} \geq \frac{\|q-q'\| - \|\pi_q^{\perp}(q-q')\|}{\|q-q'\|} \\
&\geq 1 - \frac{\|q-q'\|}{2\tau(M)} \text{ by (A.2)} \\
&\geq 1 - \frac{\|p-q\|}{\tau(M)} \text{ by the triangle inequality.}
\end{aligned}$$

Therefore, $3\|p-q\|/\tau(M) > 1$, i.e. $\|p-q\| > \tau(M)/3$ and $\tilde{\pi}_p$ is injective on $\mathcal{B}_M(p, \tau(M)/3)$. To conclude that $\tilde{\pi}_p$ is a diffeomorphism, it suffices to show that its differential is always invertible. As $\tilde{\pi}_p$ is an affine application, the differential $d_q\tilde{\pi}_p$ is equal to π_p. Therefore, the Jacobian $J\tilde{\pi}_p(q)$ of the function $\tilde{\pi}_p : M \to T_pM$ in q is given by the determinant of the projection π_p restricted to T_qM. In particular, it is larger than the smallest singular value of $\pi_p \circ \pi_q$ to the power d, which is larger than

$$(1 - \angle(T_pM, T_qM))^d \geq \left(1 - 2\frac{\|p-q\|}{\tau(M)}\right)^d \geq \left(\frac{1}{3}\right)^d,$$

thanks to Lemma A.1 and using that $\|p-q\| \leq \tau(M)/3$. In particular, the Jacobian is positive, and $\tilde{\pi}_p$ is a diffeomorphism from $\mathcal{B}_M(p, \tau(M)/3)$ to its image. The second statement of Point 2 is stated in [AL19, Lemma A.2]. The last statement is a consequence of the two first, using that if $\|x-p\| \leq \tau(M)/4$, then $8\|\tilde{\pi}_p(x) - p\|/7 \leq \tau(M)/3$.

□

Note that Point 2 was already proven in [ACLZ17, Lemma 5], but with a slightly worse constant of $\tau(M)/12$. We end this section on preliminary geometric results by stating two lemmas on the properties of convex hull built on manifolds.

Lemma A.4. *Let $p \in M$, $\sigma \subset \mathcal{B}_M(p, \tau(M)/4)$ and $\tilde{\sigma} = \tilde{\pi}_p(\sigma)$. Then,*

$$r(\tilde{\sigma}) \leq r(\sigma) \leq r(\tilde{\sigma})\left(1 + 6\frac{r(\tilde{\sigma})}{\tau(M)}\right). \tag{A.4}$$

Proof. As the projection is 1-Lipschitz, it is clear that $r(\tilde{\sigma}) \leq r(\sigma)$. Let us prove the other inequality. Let $\sigma = \{x_0, \ldots, x_k\}$, $\tilde{\sigma} = \{\tilde{x}_0, \ldots, \tilde{x}_k\}$ and fix $0 \leq i \leq k$. As $x_i \in \mathcal{B}_M(p, \tau(M)/4)$, we have by (A.3)

$$\|x_i - p\| \leq \frac{8}{7}\|\tilde{x}_i - p\| \leq \frac{16}{7}r(\tilde{\sigma}). \tag{A.5}$$

Let $\tilde{z}$ be the center of the minimum enclosing ball of $\tilde{\sigma}$. Write $\tilde{z} = \sum_{j=0}^{k} \lambda_j \tilde{x}_j$ and let $z = \sum_{j=0}^{k} \lambda_j x_j \in \mathrm{Conv}(\sigma)$. Then, we have

$$\begin{aligned}
\|z - x_i\| &\leq \|z - \tilde{z}\| + \|\tilde{z} - \tilde{x}_i\| + \|\tilde{x}_i - x_i\| \\
&\leq \sum_{j=0}^{k} \lambda_j \|x_j - \tilde{x}_j\| + r(\tilde{\sigma}) + \frac{\|x_i - p\|^2}{2\tau(M)} \text{ using (A.2)} \\
&\leq \sum_{j=0}^{k} \lambda_j \frac{\|x_j - p\|^2}{2\tau(M)} + r(\tilde{\sigma}) + \frac{128}{49}\frac{r(\tilde{\sigma})^2}{\tau(M)} \text{ using (A.2) and (A.5)} \\
&\leq r(\tilde{\sigma}) + \frac{256}{49}\frac{r(\tilde{\sigma})^2}{\tau(M)} \leq r(\tilde{\sigma}) + 6\frac{r(\tilde{\sigma})^2}{\tau(M)} \text{ using (A.5)}.
\end{aligned}$$

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

Lemma A.5. *Let $\sigma \subset M$ with $r(\sigma) < \tau(M)$ and $p \in M$ with $p \in \pi_M(\mathrm{Conv}(\sigma))$. Then,*

$$\sigma \subset \mathcal{B}\left(p, 2r(\sigma)\left(1 + \frac{r(\sigma)}{2\tau(M)}\right)\right). \tag{A.6}$$

Proof. Let $y \in \mathrm{Conv}(\sigma)$ with $\pi_M(y) = p$ and let $q \in \sigma$. One has $\|q - p\| \leq \|q - y\| + \|y - p\| \leq 2r(\sigma) + \frac{r(\sigma)^2}{\tau(M)}$ by using Lemma 3.3. □

B Proofs of Section 3

Delaunay triangulations will be at the core of the proof of Proposition 3.5 and we therefore need some preliminary definitions. A finite set will be called a *simplex* in the following, and a k-simplex is a set of cardinality $k + 1$. The circumball of a d-simplex σ in $\mathbb{R}^d$ is defined as the unique ball having the simplex σ on its boundary. It exists as long as σ does not lie on a hyperplane of $\mathbb{R}^d$. The radius of the circumball σ is called the circumradius of σ and is denoted by $\mathrm{circ}(\sigma)$. Note that in particular $\mathrm{circ}(\sigma) \geq r(\sigma)$.

A triangulation T of a finite set $A \subset \mathbb{R}^d$ is a set of d-simplices such that

1. $\bigcup_{\sigma \in T} \sigma = A$,
2. for $\sigma \neq \sigma' \in T$, the interior of $\mathrm{Conv}(\sigma)$ does not intersect the interior of $\mathrm{Conv}(\sigma')$,
3. $\bigcup_{\sigma \in T} \mathrm{Conv}(\sigma) = \mathrm{Conv}(A)$ and
4. for every $\sigma \in T$, $\mathrm{Conv}(\sigma)$ intersects A only at points of the simplex σ.

Given a finite set $A \subset \mathbb{R}^d$, a Delaunay triangulation of A is a triangulation of A such that the interior of every circumball of a simplex of the triangulation does not contain any point of A. Such a triangulation exists as long as A does not lie on a hyperplane of $\mathbb{R}^d$. It may however not be unique.

B.1 Proof of Proposition 3.5

We first show a weak version of Proposition 3.5:

Lemma B.1. *Let $M \in \mathcal{C}^2_d$ be a d-dimensional manifold and let $A \subset M$ be a finite set. If $t^*(A) \leq \tau(M)/36$, then*

$$t^*(A) \leq \varepsilon(A)\left(1 + C_1 \frac{\varepsilon(A)}{\tau(M)}\right) \quad \text{and} \tag{B.1}$$

$$t^*(A) \geq \varepsilon(A)\left(1 - C_2 \frac{\varepsilon(A)}{\tau(M)}\right). \tag{B.2}$$

Proof of inequality (B.1). For $p \in M$, define

$$t^*(p, A) := \inf\{t < \tau(M),\ p \in \pi_M(\operatorname{Conv}_t(A))\}. \tag{B.3}$$

We have $t^*(A) = \sup_{p \in M} t^*(p, A) \leq \tau(M)/36$ by assumption. Let $p \in M$ be such that $t^*(p, A) = t^*(A)$. Let $\sigma(p)$ be a simplex of A such that $p \in \pi_M(\operatorname{Conv}(\sigma(p))$, with $r(\sigma(p)) = t^*(p, A)$. Write $\tilde{\sigma}(p)$ for $\tilde{\pi}_p(\sigma(p))$. Also, let $\tilde{A}_p = \tilde{\pi}_p(A \cap \mathcal{B}(p, \tau(M)/4))$.

Lemma B.2. *The set $\tilde{A}_p$ does not lie on a hyperplane of $\tilde{T}_p M$.*

Proof. Assume that it lies on some hyperplane H of $\tilde{T}_p M$ and consider a point $\tilde{q} \in \tilde{T}_p M$ nearby p with $\tilde{q} - p$ orthogonal to H. Then, by Point 2 in Lemma A.3, there exists $q \in \mathcal{B}_M(p, \tau(M)/4)$ with $\tilde{\pi}_p(q) = \tilde{q}$, and q belongs to $\pi_M(\operatorname{Conv}(\sigma(q))$ for some simplex $\sigma(q)$ of radius smaller than $\sigma(p)$. Therefore, if $t^*(A) = r(\sigma(p)) \leq \tau(M)/9$, then by Lemma A.5,

$$\begin{aligned}
\sigma(q) &\subset \mathcal{B}_M\left(q, 2r(\sigma(q))\left(1 + \frac{r(\sigma(q))}{2\tau(M)}\right)\right) \\
&\subset \mathcal{B}_M(q, 19\tau(M)/81) \\
&\subset \mathcal{B}_M(p, \|p - q\| + 19\tau(M)/81) \subset \mathcal{B}_M(p, \tau(M)/4)
\end{aligned}$$

by choosing $\tilde{q}$ close enough to p and using (A.3). Hence, we have $\tilde{\pi}_p(\sigma(q)) \subset \tilde{A}_p \subset H$. Let $y \in \operatorname{Conv}(\sigma(q))$ be such that $\pi_M(y) = q$. Then, $\tilde{\pi}_p(y) \in H$. Therefore, by Pythagoras' theorem,

$$\begin{aligned}
\|y - p\|^2 &= \|y - \tilde{q}\|^2 - \|p - \tilde{q}\|^2 \\
&\leq (\|y - q\| + \|q - \tilde{q}\|)^2 - \|p - \tilde{q}\|^2 \\
&\leq \|y - q\|^2 + \|q - \tilde{q}\|^2 + 2\|y - q\|\|q - \tilde{q}\| - \|p - \tilde{q}\|^2.
\end{aligned} \tag{B.4}$$

By (A.2) and Lemma A.3, we have $\|q-\tilde{q}\| \leq \|p-q\|^2/(2\tau(M)) \leq 64\|p-\tilde{q}\|^2/(98\tau(M))$, as long as $\|p-\tilde{q}\|$ is sufficiently small. Also, by Lemma 3.3, $\|y-q\| \leq r(\sigma(q))^2/\tau(M) \leq \tau(M)/(36)^2$. Therefore, from (B.4), we obtain that $\|y-p\| < \|y-q\|$ if $\|p-\tilde{q}\|$ is sufficiently small. This is a contradiction with having $\pi_M(y) = q$. Therefore, $\tilde{A}_p$ does not lie on H. □

Therefore, there exists a Delaunay triangulation of $\tilde{A}_p$, which we will consider in the following. If $t^*(p,A) \leq \tau(M)/9$, then $\sigma(p) \subset \mathcal{B}(p,\tau(M)/4)$ according to Lemma A.5. Therefore, using Point 1 in Lemma A.3, we see that $p \in \tilde{\pi}_p(\mathrm{Conv}(\sigma(p))) \subset \mathrm{Conv}(\tilde{A}_p)$ and that there exists a simplex $\tilde{\sigma}_0$ in a Delaunay triangulation of $\tilde{A}_p$ with $p \in \mathrm{Conv}(\tilde{\sigma}_0)$. We denote by σ_0 be the corresponding simplex in A.

Lemma B.3. *Assume that $p \in M$ satisfies $t^*(p,A) \leq \tau(M)/36$ and that there exists $\tilde{y} \in \tilde{T}_pM$ with $\|p-\tilde{y}\| \leq 3t^*(p,A)$. Then, there exists $y \in M$ with $\tilde{\pi}_p(y) = \tilde{y}$ and $d(y,A) \geq d(\tilde{y},\tilde{A}_p)$.*

Before proving Lemma B.3, let us finish the proof. Let $\tilde{z}$ be the center of the smallest enclosing ball of $\tilde{\sigma}(p)$ and $\tilde{w}$ be the center of the circumsphere of $\tilde{\sigma}_0$. We apply Lemma B.3 on a certain $\tilde{y}$, which is built in a different way, depending on whether $\tilde{w}$ and $\tilde{z}$ are close or not.

- **Case 1:** Assume that $\|\tilde{z}-\tilde{w}\| \leq 2r(\tilde{\sigma}(p))$. Then, we choose $\tilde{y} := \tilde{w}$. Indeed, we have:
  - $\|p-\tilde{w}\| \leq \|p-\tilde{z}\| + \|\tilde{z}-\tilde{w}\| \leq r(\tilde{\sigma}(p)) + 2r(\tilde{\sigma}(p)) = 3r(\tilde{\sigma}(p)) \leq 3t^*(p,A)$. The second inequality holds as $p \in \mathrm{Conv}(\tilde{\sigma}(p)) \subset \mathcal{B}(\tilde{z}, r(\tilde{\sigma}(p)))$.
  - $d(\tilde{w},\tilde{A}_p) = \mathrm{circ}(\tilde{\sigma}_0) \geq r(\tilde{\sigma}_0)$ as $\tilde{\sigma}_0$ is in the Delaunay triangulation.

  Therefore, one can apply Lemma B.3 to $\tilde{w}$: one has $\varepsilon(A) \geq d(w,A) \geq d(\tilde{w},\tilde{A}_p) \geq r(\tilde{\sigma}_0)$ for some $w \in M$. Also, there exists an element $y \in \mathrm{Conv}(\sigma_0)$ with $\tilde{\pi}_p(y) = p$ by construction. As $r(\sigma_0) \leq \tau(M)/4$, we have $\pi_M(y) = p$, according to Point 1 in Lemma A.3. This implies that $t^*(p,A) \leq r(\sigma_0)$. Therefore, according to Lemma A.4, as $\sigma_0 \subset \mathcal{B}_M(p,\tau(M)/4)$ by construction,

$$t^*(p,A) \leq r(\sigma_0) \leq r(\tilde{\sigma}_0)\left(1+6\frac{r(\tilde{\sigma}_0)}{\tau(M)}\right) \leq \varepsilon(A)\left(1+6\frac{\varepsilon(A)}{\tau(M)}\right).$$

- **Case 2:** Assume that $\|\tilde{z}-\tilde{w}\| > 2r(\tilde{\sigma}(p))$. Consider $\tilde{y} = \tilde{z} + 2r(\tilde{\sigma}(p))\frac{\tilde{w}-\tilde{z}}{\|\tilde{w}-\tilde{z}\|}$. Then, we have:
  - $\|p-\tilde{y}\| \leq \|p-\tilde{z}\| + \|\tilde{z}-\tilde{y}\| \leq r(\tilde{\sigma}(p)) + 2r(\tilde{\sigma}(p)) = 3r(\tilde{\sigma}(p)) \leq 3t^*(p,A)$.
  - $\|\tilde{y}-\tilde{w}\| = \|\tilde{z}-\tilde{w}\| - 2r(\tilde{\sigma}(p)) \leq \|\tilde{z}-p\| + \|p-\tilde{w}\| - 2r(\tilde{\sigma}(p)) \leq \|p-\tilde{w}\| - r(\tilde{\sigma}(p))$. As p is in the circumball of $\tilde{\sigma}_0$, $\|\tilde{y}-\tilde{w}\| \leq \|p-\tilde{w}\| \leq \mathrm{circ}(\tilde{\sigma}_0)$, i.e. $\tilde{y}$ is also in the circumball of $\tilde{\sigma}_0$. Therefore, letting $\mathcal{S}$ be the circumsphere of $\tilde{\sigma}_0$,

$$\begin{aligned} d(\tilde{y},\tilde{A}_p) &\geq d(\tilde{y},\mathcal{S}) = \mathrm{circ}(\tilde{\sigma}_0) - \|\tilde{y}-\tilde{w}\| \\ &\geq \mathrm{circ}(\tilde{\sigma}_0) - \|p-\tilde{w}\| + r(\tilde{\sigma}(p)) \geq r(\tilde{\sigma}(p)). \end{aligned}$$

Likewise the first case, one can apply Lemma B.3 to $\tilde{y}$ and obtain $\varepsilon(A) \geq r(\tilde{\sigma}(p))$. Therefore, using Lemma A.4,

$$t^*(p, A) = r(\sigma(p)) \leq r(\tilde{\sigma}(p)) \left(1 + 6\frac{r(\tilde{\sigma}(p))}{\tau(M)}\right) \leq \varepsilon(A) \left(1 + 6\frac{\varepsilon(A)}{\tau(M)}\right).$$

We therefore have shown $t^*(A) = t^*(p, A) \leq \varepsilon(A) \left(1 + C_1 \frac{\varepsilon(A)}{\tau(M)}\right)$ with $C_1 = 6$. □

Proof of Lemma B.3. According to Point 2 in Lemma A.3, if $3t^*(p, A) \leq 7\tau(M)/32$, then there exists $y \in \mathcal{B}_M(p, \tau(M)/4)$ with $\tilde{\pi}_p(y) = \tilde{y}$. Then, $d(\tilde{y}, \tilde{A}_p) \leq d(y, A \cap \mathcal{B}(p, \tau(M)/4))$ as the projection is 1-Lispchitz. To conclude, it suffices to show that $d(y, A \cap \mathcal{B}(p, \tau(M)/4)) = d(y, A)$. If this is not the case, then there exists $a \in A$ with $\|p - a\| > \tau(M)/4$ and $\|y - a\| \leq d(y, A \cap \mathcal{B}(p, \tau(M)/4))$, so that

$$\|y - p\| + d(y, A \cap \mathcal{B}(p, \tau(M)/4)) \geq \|y - p\| + \|y - a\| \geq \|p - a\| > \tau(M)/4.$$

Let $\sigma(p)$ be a simplex of A with $r(\sigma(p)) = t^*(p, A)$ and $p \in \pi_M(\mathrm{Conv}(\sigma(p))$. By Lemma A.5, $\sigma(p) \subset \mathcal{B}(p, \tau(M)/4)$. Therefore, for $x \in \sigma(p)$, we have

$$d(y, A \cap \mathcal{B}(p, \tau(M)/4)) \leq \|y - x\| \leq \|y - p\| + \|p - x\|.$$

According to Lemma A.5, one has $\|p - x\| \leq 2t^*(p, A)\,(1 + t^*(p, A)/(2\tau(M)))$. Also, according to (A.3), $\|y - p\| \leq 8\|\tilde{y} - p\|/7$. Therefore,

$$\begin{aligned}
\|y - p\| + d(y, A \cap \mathcal{B}(p, \tau(M)/4)) &\leq \frac{16}{7}\|\tilde{y} - p\| + \|p - x\| \\
&\leq \frac{16}{7} 3t^*(p, A) + 2t^*(p, A) \left(1 + \frac{t^*(p, A)}{2\tau(M)}\right) \\
&\leq \tau(M)/4 \text{ if } t^*(p, A) \leq \frac{\tau(M)}{36},
\end{aligned}$$

which concludes the proof. □

Proof of inequality (B.2). Let $p \in M$. There exist a simplex $\sigma(p)$ and $x \in \mathrm{Conv}(\sigma(p))$ with $\pi_M(x) = p$ and $r(\sigma(p)) \leq t^*(A)$. By Lemma 1 in [ALS13], we have $d(x, \sigma(p)) \leq r(\sigma(p))$, i.e. there exists $q \in \sigma(p)$ with $\|x - q\| \leq r(\sigma(p))$. Then,

$$\begin{aligned}
d(p, A) \leq \|p - q\| &\leq \|p - x\| + \|x - q\| \\
&\leq \frac{t^*(A)^2}{\tau(M)} + t^*(A) \text{ by Lemma 3.3.}
\end{aligned}$$

By taking the supremum over $p \in M$ in, we obtain $\varepsilon(A) \leq t^*(A)\left(1 + \frac{t^*(A)}{\tau(M)}\right)$. In particular, $\varepsilon(A) \leq 2t^*(A)$, and by using (B.1), we obtain that, if $t^*(A) \leq \tau(M)/36$,

$$\varepsilon(A) \leq t^*(A)\left(1 + \frac{\varepsilon(A)\left(1 + C_1\frac{\varepsilon(A)}{\tau(M)}\right)}{\tau(M)}\right) \leq t^*(A)\left(1 + \frac{\varepsilon(A)\left(1 + (2C_1)/36\right)}{\tau(M)}\right),$$

so that $\varepsilon(A)\left(1 - (1 + (2C_1)/36)\,\frac{\varepsilon(A)}{\tau(M)}\right) \leq t^*(A)$ as long as $t^*(A) \leq \tau(M)/36$. □

Proof of Proposition 3.5. To prove Proposition 3.5, by using Lemma B.1, it suffices to show that there exists two absolute constants c_0, c_1 for which

$$t^*(A) \leq c_0\varepsilon(A) \text{ if } \varepsilon(A) \leq c_1\tau(M). \tag{B.5}$$

Lemma B.4. *Let $A \subset \mathbb{R}^d$ be a finite set. If $d_H(\mathcal{B}(0,1)|A) \leq 1$, then $0 \in \mathrm{Conv}(A)$.*

Proof. We prove the contrapositive. If $0 \notin \mathrm{Conv}(A)$, then there exists an half space which contains A. Let x be the unit vector orthogonal to this halfspace. Then, $d(x, A) > 1$. □

Let $p \in M$ and let $\tilde{y} \in \mathcal{B}_{\tilde{T}_pM}(p, \varepsilon(A))$. If $\varepsilon(A) \leq 7\tau(M)/24$, then there exists $y \in \mathcal{B}_M(p, 8\varepsilon(A)/7)$ with $\tilde{\pi}_p(y) = \tilde{y}$ according to Point 2 in Lemma A.3. By assumption, there exists $a \in A$ with $\|y - a\| \leq \varepsilon(A)$, and this point a is in $\mathcal{B}(p, 15\varepsilon(A)/7)$. Therefore, as $\tilde{\pi}_p$ is 1-Lipschitz,

$$d_H(\mathcal{B}_{\tilde{T}_pM}(p, \varepsilon(A))|\tilde{\pi}_p(A \cap \mathcal{B}(p, 15\varepsilon(A)/7))) \leq \varepsilon(A). \tag{B.6}$$

By Lemma B.4, this implies that

$$p \in \mathrm{Conv}(\tilde{\pi}_p(A \cap \mathcal{B}(p, 15\varepsilon(A)/7))) = \tilde{\pi}_p(\mathrm{Conv}(A \cap \mathcal{B}(p, 15\varepsilon(A)/7))).$$

If $15\varepsilon(A)/7 < \tau(M)$, then there is $x \in \mathrm{Conv}(A \cap \mathcal{B}(p, 15\varepsilon(A)/7))$ with $\pi_M(x) = p$ according to Point 1 in Lemma A.3 and Lemma 3.3. As this holds for any $p \in M$, we have

$$t^*(A) \leq \sup_{p \in M} r(A \cap \mathcal{B}(p, 15\varepsilon(A)/7)) \leq \frac{15\varepsilon(A)}{7}, \tag{B.7}$$

as long as $\varepsilon(A) < 7\tau(M)/24$, thus showing (B.5) and concluding the proof of Proposition 3.5. □

B.2 Proof of Theorem 3.7

We first state a lemma which shows that the t-convex hull is stable under small perturbations with respect to the Hausdorff distance.

Lemma B.5. *Let $t, \gamma > 0$ and $A, B \subset \mathbb{R}^D$ with $d_H(A, B) \leq \gamma$. Then,*

$$d_H(\operatorname{Conv}_t(B) | \operatorname{Conv}_{t+\gamma}(A)) \leq \gamma. \tag{B.8}$$

Proof. Let $\sigma \subset B$ be a non-empty set with $r(\sigma) \leq t$. Let $\xi = \{x \in A,\ d(x, \sigma) \leq \gamma\}$. By assumption, ξ is non-empty and $d_H(\sigma, \xi) \leq \gamma$. One has $r(\xi) \leq t + \gamma$ (see [ALS13, Lemma 16]) and $d_H(\operatorname{Conv}(\sigma) | \operatorname{Conv}(\xi)) \leq d_H(\sigma | \xi) \leq \gamma$. This implies the conclusion. □

Let $A \subset M$ and $B \subset \mathbb{R}^D$ with $d_H(A, B) \leq \gamma$. Then, if $t^*(A) < t + \gamma < \tau(M)$,

$$\begin{aligned} d_H(\operatorname{Conv}_t(B) | M) &\leq d_H(\operatorname{Conv}_t(B) | \operatorname{Conv}_{t+\gamma}(A)) + d_H(\operatorname{Conv}_{t+\gamma}(A), M) \text{ by (2.4)} \\ &\leq \gamma + \frac{(t+\gamma)^2}{\tau(M)} \text{ by Lemma B.5 and (3.4).} \end{aligned} \tag{B.9}$$

Let $P \in \mathcal{P}_{d, \tau_{\min}, f_{\min}, \gamma}$ be supported on some manifold M and let $\mathbb{X}_n$ be a n-sample of law P, with $\mathbb{Y}_n$ the corresponding sample on M. Then, for $0 \leq t < \tau(M) - \gamma$,

$$\begin{aligned} \mathbb{E} d_H(\operatorname{Conv}_t(\mathbb{Y}_n), M) &= \mathbb{E} d_H(\operatorname{Conv}_t(\mathbb{X}_n), M) \mathbf{1}\{t + \gamma > t^*(\mathbb{Y}_n)\} \\ &\quad + \mathbb{E} d_H(\operatorname{Conv}_t(\mathbb{X}_n), M) \mathbf{1}\{t + \gamma \leq t^*(\mathbb{Y}_n)\} \\ &\leq \gamma + \frac{(t+\gamma)^2}{\tau(M)} + (\operatorname{diam}(M) + \gamma) \mathbb{P}(t^*(\mathbb{Y}_n) \geq t + \gamma). \end{aligned}$$

By Proposition 3.5, if $\varepsilon(\mathbb{Y}_n) \leq C_0 \tau(M)$, then $t^*(\mathbb{Y}_n) \geq t + \gamma$ implies that

$$\varepsilon(\mathbb{Y}_n) \geq (t + \gamma) \left(1 + C_1 \frac{\varepsilon(\mathbb{Y}_n)}{\tau(M)}\right)^{-1} \geq C_2 (t + \gamma)$$

for some absolute constant C_2. Therefore, $t^*(\mathbb{Y}_n) \geq t + \gamma$ implies

$$\varepsilon(\mathbb{Y}_n) \geq \min\left(C_0 \tau(M), C_2(t + \gamma)\right) = C_2(t + \gamma) \tag{B.10}$$

if $t + \gamma \leq C_0 \tau(M) / C_2$. By using Proposition 3.6, and by noting that $\operatorname{diam}(M)$ is bounded by a constant depending on $d, f_{\min}, \tau_{\min}$ (see [AL18, Lemma 2]), we obtain that, if $t + \gamma \leq C_0 \tau(M) / C_2$,

$$\mathbb{E} d_H(\operatorname{Conv}_t(\mathbb{Y}_n), M) \leq \gamma + \frac{(t+\gamma)^2}{\tau(M)} + c_{d, \tau_{\min}, f_{\min}} \frac{1}{(t+\gamma)^d} \exp(-C_{d, f_{\min}} n (t + \gamma)^d). \tag{B.11}$$

In particular, we obtain the conclusion by letting $t = C_{d, \tau_{\min}, f_{\min}} (\log n / n)^{1/d}$ for $C_{d, \tau_{\min}, f_{\min}}$ sufficiently large, if $\gamma \leq \eta (\log n / n)^{2/d}$.

C Proofs of Section 4

C.1 Proof of Proposition 4.3

Let $P \in \mathcal{P}_{d,\tau_{\min},f_{\min},f_{\max}}$ be a probability distribution with support M and let $\mathbb{X}_n$ be a n-sample of law P. We will use repeatedly in the proof the fact that there exist constants $c_d, C_d > 0$ such that, if $t \leq \tau(M)/4$, then $c_d f_{\min} t^d \leq P(B) \leq C_d f_{\max} t^d$ for all balls B of radius t centered at points of M (see [Aam17, Lemma III.23]).

Lemma C.1. *Assume that $t \leq t_{d,\tau_{\min},f_{\max}}$. Then, there exists a partition $\mathcal{C} = \{U_1, \dots, U_K\}$ of M into K measurable parts such that:*

1. *for $k = 1, \dots, K$, U_k contains a ball B_k of radius $2t$,*
2. *for $k = 1, \dots, K$, $P(U_k) = 1/K$,*
3. *we have $K \asymp_{d,\tau_{\min},f_{\max}} t^{-d}$.*

Proof. If $t \leq \tau(M)/8$, then $P(B) \leq C_d f_{\max} t^d$ for any ball B of radius $2t$. Assume that t is small enough so that $C_d f_{\max} t^d \leq 1/2$ and let K be the largest integer such that $1/K \geq C_d f_{\max} t^d$, so that $1/(2C_d f_{\max} t^d) \leq K \leq 1/(C_d f_{\max} t^d)$. Build $\mathcal{C}$ in the following way. Start with an union of K disjoint balls B_k of radius $2t$, for $k = 1, \dots, K$, choose V_k any measurable set in $M \setminus \bigcup_{k=1}^K B_k$ with $P(V_k) = 1/K - P(B_k) \geq 0$ and let $U_k = B_k \cup V_k$. The set $M \setminus \bigcup_{k=1}^K U_k$ is of P-measure null, so that by adding it to U_1 for instance, we obtain a partition following the required properties. Note that we used the fact that for any $A \subset M$ and $0 \leq p \leq P(A)$, there exists a subset $V \subset A$ with $P(V) = p$: this holds as P is absolutely continuous with respect to the volume measure on M. □

We fix such a partition in the following. For $V \subset M$, let N_V be the number of points of $\mathbb{X}_n$ in V and write N_k for N_{U_k}. Denote by B'_k the ball sharing its center with B_k, of radius t and define E_k the event

$$\begin{aligned} &(N_k = 2 \text{ and } N_{B'_k} = 2) \Rightarrow r(\mathbb{X}_n \cap U_k) < \lambda t \\ \equiv\ & N_k \neq 2 \text{ or } (N_k = 2 \text{ and } (N_{B'_k} < 2 \text{ or } (N_{B'_k} = 2 \text{ and } r(\mathbb{X}_n \cap U_k) < \lambda t))) \\ \equiv\ & N_k \neq 2 \text{ or } F_k. \end{aligned} \tag{C.1}$$

Lemma C.2. *If $h(t, \mathbb{X}_n) < \lambda t$, then E_k is satisfied for $k = 1, \dots, K$.*

Proof. Let $\sigma = \mathbb{X}_n \cap U_k$. If $N_k = 2$ and $N_{B'_k} = 2$, then both points of σ are in B'_k and one has $r(\sigma) \leq t$. Therefore, $d_H(\mathrm{Conv}(\sigma)|\mathbb{X}_n) < \lambda t$. Let X_e be the middle of the two points composing σ. The smallest enclosing ball of σ is of radius smaller than t, and is therefore included in B_k (which is of radius $2t$). As $N_{B_k} = 2$, one has $d(X_e, \mathbb{X}_n) = d(X_e, \sigma) = r(\sigma)$. Therefore, we have $r(\sigma) \leq d_H(\mathrm{Conv}(\sigma)|\mathbb{X}_n) < \lambda t$ and E_k is satisfied. □

We therefore obtain the bound

$$\begin{aligned}\mathbb{P}(h(t,\mathbb{X}_n)<\lambda t) &\le \mathbb{P}(\forall k=1,\dots,K,\ E_k)\\ &= \mathbb{E}\left[\mathbb{P}(\forall k=1,\dots,K,\ E_k|(N_k)_{k=1,\dots,K})\right]\\ &\le \mathbb{E}\left[\prod_{k=1}^{K}(\mathbf{1}\{N_k\neq 2\}+\mathbb{P}(F_k|N_k=2)\mathbf{1}\{N_k=2\})\right]\\ &\le \mathbb{E}\left[\prod_{k=1}^{K}(1-(1-\mathbb{P}(F_k|N_k=2))\mathbf{1}\{N_k=2\})\right].\end{aligned}$$

Lemma C.3. *There exists a positive constant C_0 (depending on $\lambda, d, f_{\min}, f_{\max}$) such that*

$$\mathbb{P}(F_k|N_k=2)\le e^{-C_0} \text{ for } k=1,\dots,K.$$

Proof. Let Y_1, Y_2 be two independent random variables sampled according to P, conditioned on being in B'_k. Then,

$$\begin{aligned}\mathbb{P}(F_k|N_k=2) &= \mathbb{P}(N_{B'_k}<2|N_k=2)+\mathbb{P}(N_{B'_k}=2 \text{ and } r(\mathbb{X}_n\cap U_k)<\lambda t|N_k=2)\\ &= 1-\mathbb{P}(N_{B'_k}=2 \text{ and } r(\mathbb{X}_n\cap U_k)\ge\lambda t|N_k=2)\\ &= 1-\mathbb{P}(N_{B'_k}=2|N_k=2)\mathbb{P}(r(\mathbb{X}_n\cap B'_k)\ge\lambda t|N_{B'_k}=2)\\ &= 1-\left(\frac{P(B'_k)}{P(U_k)}\right)^2\mathbb{P}(r(\{Y_1,Y_2\})\ge\lambda t)\\ &\le 1-\left(KC_d f_{\min}t^d\right)^2\mathbb{P}(\|Y_1-Y_2\|\ge 2\lambda t) \text{ using [Aam17, Lemma III.23]}\\ &\le 1-C_1\mathbb{P}(\|Y_1-Y_2\|\ge 2\lambda t) \text{ using Lemma C.1.}\end{aligned}$$

Let x_1, x_2 be two opposite points on B'_k. If $\|x_i-Y_i\|\le(1-\lambda)t$ for $i=1,2$, then $\|Y_1-Y_2\|\ge 2\lambda t$. Also, there exists a ball W_i of radius $(1-\lambda)t/2$ in $\mathcal{B}(x_i,(1-\lambda)t)\cap B'_k$. Therefore,

$$\mathbb{P}(\|Y_1-Y_2\|\ge 2\lambda t)\ge\left(\frac{P(W_i)}{P(B'_k)}\right)^2\ge\left(\frac{C_d f_{\min}\left(\frac{(1-\lambda)t}{2}\right)^d}{C_d f_{\max}t^d}\right)^2=C_2,$$

where we used [Aam17, Lemma III.23]. Thus, the lemma holds with $C_0:=-\log(1-C_1C_2)$. □

We finally obtain

$$\mathbb{P}(h(t,\mathbb{X}_n)<\lambda t)\le\mathbb{E}\left[\exp\left(-C_0\sum_{k=1}^{K}\mathbf{1}\{N_k=2\}\right)\right]. \tag{C.2}$$

We use the following theorem to estimate this quantity (see [LL14]):

Proposition C.4. *Let $Z_1, \dots, Z_K$ be Bernoulli random variables. Let $0 < l < L < K$ be positive integers. Then,*

$$\mathbb{P}\left(\sum_{k=1}^{K} Z_k \geq L\right) \leq \frac{1}{\binom{L}{l}} \sum_{\substack{A \subset \{1,\dots,K\} \\ |A|=l}} \mathbb{E}\left[\prod_{i \in A} Z_i\right], \tag{C.3}$$

where $|A|$ denotes the cardinality of a set A.

For $k = 1, \dots, K$ and $n > 0$, let $Z_k := \mathbf{1}\{N_k \neq 2\}$, $I_k(n) := \mathbb{E}\left[\prod_{l=1}^{k} Z_l\right]$ and

$$p := \mathbb{P}(N_k = 2) = \binom{n}{2} K^{-2} \left(1 - \frac{1}{K}\right)^{n-2}. \tag{C.4}$$

Assume that $K \geq 17$ (this can be ensured by taking t small enough according to Lemma C.1). Then,

$$p \leq \frac{\frac{1}{2}\left(\frac{n}{K}\right)^2 \exp(-n/K)}{\left(1 - \frac{1}{17}\right)^2} \leq 1/3.$$

One has, for $k \geq 1$ and $n \geq 2$,

$$\begin{aligned} I_k(n) &= \mathbb{P}(N_1 \neq 2, \dots, N_{k-1} \neq 2) - \mathbb{P}(N_1 \neq 2, \dots, N_{k-1} \neq 2, N_k = 2) \\ &= I_{k-1}(n) - \mathbb{P}(N_1 \neq 2, \dots, N_{k-1} \neq 2 | N_k = 2)p \\ &= I_{k-1}(n) - I_{k-1}(n-2)p. \end{aligned}$$

Let, for $k \geq 1$ and $n \geq 2$,

$$R_k(n) := \frac{I_k(n)}{I_{k-1}(n)} \text{ and } S_k(n) := \frac{I_k(n-2)}{I_k(n)}, \tag{C.5}$$

so that $R_k(n) = 1 - S_{k-1}(n)p$. One has

$$\begin{aligned} I_k(n) &= \mathbb{P}(N_1 \neq 2, \dots, N_k \neq 2 \text{ and } X_1 \notin \bigcup_{l \leq k} U_l) + \mathbb{P}(N_1 \neq 2, \dots, N_k \neq 2 \text{ and } X_1 \in \bigcup_{l \leq k} U_l) \\ &= I_k(n-1)\left(1 - \frac{k}{K}\right) + \mathbb{P}(N_1 \neq 2, \dots, N_k \neq 2 \text{ and } X_1 \in \bigcup_{l \leq k} U_l), \end{aligned}$$

so that

$$\left(1 - \frac{k}{K}\right) I_k(n-1) \leq I_k(n) \leq I_k(n-1) + I_{k-1}(n-1). \tag{C.6}$$

Iterating this equation, we obtain

$$\left(1 - \frac{k}{K}\right)^2 I_k(n-2) \leq I_k(n) \leq I_k(n-2) + 2I_{k-1}(n-2) + I_{k-2}(n-2).$$

Therefore,

$$\left(1-\frac{k}{K}\right)^2 \le S_k(n)^{-1} \le 1 + 2R_k(n-2)^{-1} + R_k(n-2)^{-1}R_{k-1}(n-2)^{-1}. \tag{C.7}$$

Assume that $2 \le k \le K(1-(3/2)\sqrt{p})$ (with $1-(3/2)\sqrt{p} > 0$ for $p \le 1/3$). Then,

$$R_{k-1}(n) = 1 - S_{k-2}(n)p \ge 1 - \frac{p}{\left(1-\frac{k-2}{K}\right)^2} \ge 1 - \left(\frac{2}{3}\right)^2 > 0.$$

Therefore, by (C.7), if $3 \le k \le K(1-(3/2)\sqrt{p})$, then $S_{k-1}(n)^{-1} \le C_3$ for some absolute constant C_3 and

$$R_k(n) = 1 - S_{k-1}(n)p \le 1 - C_3^{-1}p. \tag{C.8}$$

Thus, we have, for $3 \le l \le K(1-(3/2)\sqrt{p})$,

$$I_l(n) = \prod_{k=1}^{l} R_k(n) \le \prod_{k=3}^{l} R_k(n) \le \left(1-C_3^{-1}p\right)^{l-2} \le C_4 \exp(-C_3^{-1}lp). \tag{C.9}$$

We are now ready to apply Proposition C.4 to $Z_1, \dots, Z_K$ for some integers l, K, with $3 \le l \le K(1-(3/2)\sqrt{p}) < L < K$:

$$\mathbb{P}\left(\sum_{k=1}^{K} \mathbf{1}\{N_k = 2\} \le K - L\right) = \mathbb{P}\left(\sum_{k=1}^{K} Z_k \ge L\right) \le \frac{\binom{K}{l}}{\binom{L}{l}} C_4 \exp(-C_3^{-1}lp). \tag{C.10}$$

To conclude, we use the following estimate:

Lemma C.5. *There exists an absolute constant μ such that the following holds. Let $0 < p \le 1/3$ and let $K \ge 17$ be an integer satisfying*

$$-K\mu p/\log(p) \ge 1. \tag{C.11}$$

Then there exists integers l, L such that

$$2 < K/8 \le l \le K(1-(3/2)\sqrt{p}) < L \le K(1+\mu p/\log(p)) < K$$

and

$$\frac{\binom{K}{l}}{\binom{L}{l}} \le C_5 \exp((C_3^{-1}/16)Kp), \tag{C.12}$$

for some absolute constant C_5.

Before proving Lemma C.5, let us finish the proof. Assume first that K and n are such that condition (C.11) is satisfied and choose integers l, L as in Lemma C.5 to obtain from (C.10) that

$$\begin{aligned}\mathbb{P}\left(\sum_{k=1}^{K}\mathbf{1}\{N_k=2\}\le -K\mu p/\log(p)\right) &\le C_4C_5\exp(-(C_3^{-1}/8)Kp+(C_3^{-1}/16)Kp))\\ &= C_4C_5\exp(-(C_3^{-1}/16)Kp).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\exp\left(-C_0\sum_{k=1}^{K}\mathbf{1}\{N_k=2\}\right)\right] &\le \mathbb{P}\left(\sum_{k=1}^{K}\mathbf{1}\{N_k=2\}\le -K\mu p/\log(p)\right)+\exp(C_0K\mu p/\log(p))\\ &\le C_4C_5\exp(-(C_3^{-1}/16)Kp)+\exp(C_0K\mu p/\log(p))\\ &\le C_6\exp(C_7Kp/\log(p)).\end{aligned}\tag{C.13}$$

Note that, should condition (C.11) be not satisfied, then the left-hand side of (C.13) is larger than $C_6\exp(-C_7/\mu)$. Thus, by replacing C_6 by a larger constant if necessary, the left-hand side of (C.13) is larger than 1 in this case. As the right-hand side of (C.13) is smaller than 1, we observe that (C.13) holds even if (C.11) is not satisfied. Also, for $K\ge 17$, one can easily check that $p\ge (n/2K)^2e^{-2n/K}\ge C_8(nt^d)^2\exp(-C_9nt^d)$. As the function $p\in(0,1)\mapsto p/\log(p)$ is nonincreasing, this concludes the proof.

The only remaining part is to prove Lemma C.5.

Proof of Lemma C.5. We first prove that there exists integers $2<l<L<K$ satisfying

$$\begin{aligned}&K/8\le K(1-(3/2)\sqrt{p}+\mu p/\log(p))\le l\le K(1-(3/2)\sqrt{p}) \text{ and}\\ &K(1+2\mu p/\log(p))\le L\le K(1+\mu p/\log(p))<K.\end{aligned}\tag{C.14}$$

Indeed, one has $1-(3/2)\sqrt{p}>1/8$ as $p\le 1/3$, and also, for any $\kappa>0$, $\sqrt{p}>-\kappa\mu p/\log(p)$ for $0<p\le 1/3$ if μ is sufficiently small with respect to κ. Therefore, $K/8\le K(1-(3/2)\sqrt{p}+\mu p/\log(p))$ and $K(1-(3/2)\sqrt{p})<K(1+2\mu p/\log(p))$ for μ small enough. The existence of integers L and l satisfying (C.14) is then ensured by the inequality $-K\mu p/\log(p)\ge 1$. We now fix such integers l, L.

To prove (C.12), we use the following bound which holds for any $0<k<K$ (see [Gal, Exercise 5.8]):

$$\sqrt{\frac{K}{8k(K-k)}}\exp\left(Kh(k/K)\right)\le\binom{K}{k}\le\sqrt{\frac{K}{2\pi k(K-k)}}\exp\left(Kh(k/K)\right),\tag{C.15}$$

where $h(x)=-x\ln x-(1-x)\ln(1-x)$ for $x\in(0,1)$. There exists an absolute constant c_0 such that $|h'(x)|\le -c_0\log(1-x)$ for $x\in(1/8,1)$. Therefore, as $1/8\le 1-(3/2)\sqrt{p}+\mu p/\log(p)\le$

$l/K \le 1-(3/2)\sqrt{p}$,

$$\begin{aligned} h(l/K) &= h(1-(3/2)\sqrt{p}) + (h(l/K) - h(1-(3/2)\sqrt{p})) \\ &\le h(1-(3/2)\sqrt{p}) + c_0 \log((3/2)\sqrt{p})\mu p/\log(p) \\ &\le h(1-(3/2)\sqrt{p}) + c_1 \mu p, \end{aligned}$$

as there exists $\alpha > 0$ such that $(3/2)\sqrt{p} \ge p^\alpha$ for $0 < p \le 1/3$. Therefore, using that the function $x \in (0,1) \mapsto x^{-1}h(x)$ is nonincreasing,

$$\begin{aligned} &\frac{\binom{K}{l}}{\binom{L}{l}} \le \sqrt{\frac{8K(L-l)}{2\pi L(K-l)}} \exp\left(Kh(l/K) - Lh(l/L)\right) \\ &\le \sqrt{\frac{8(1-l/L)}{2\pi(1-l/K)}} \exp\left(Kh(1-(3/2)\sqrt{p}) + c_1\mu pK - l\frac{1+2\mu p/\log(p)}{1-(3/2)\sqrt{p}} h\left(\frac{1-(3/2)\sqrt{p}}{1+2\mu p/\log(p)}\right)\right) \\ &\le \sqrt{\frac{8}{2\pi}} \exp\Bigg(K\Bigg(h(1-(3/2)\sqrt{p}) + c_1\mu p \\ &\qquad - (1-(3/2)\sqrt{p} + \mu p/\log(p)) \frac{1+2\mu p/\log(p)}{1-(3/2)\sqrt{p}} h\left(\frac{1-(3/2)\sqrt{p}}{1+2\mu p/\log(p)}\right)\Bigg)\Bigg) \\ &= \sqrt{\frac{8}{2\pi}} \exp\left(K(F_\mu(p) + c_1\mu p)\right) \end{aligned}$$

Let us bound $F_\mu(p)$. Write $F_\mu(p) = h(a) - bh(c)$, so that $F_\mu(p) = h(a)(1-b) - b(h(c)-h(a))$.

- One has, using $1-(3/2)\sqrt{p} \ge 1/8$,

$$\begin{aligned} 1-b &= \frac{1-(3/2)\sqrt{p} - (1-(3/2)\sqrt{p} + \mu p/\log(p))(1+2\mu p/\log(p))}{1-(3/2)\sqrt{p}} \\ &= \frac{-3\mu p/\log(p) - 2(\mu p/\log(p))^2 + 3\mu p^{3/2}/\log(p)}{1-(3/2)\sqrt{p}} \\ &\le -24\mu p/\log(p), \end{aligned}$$

  and also it is clear from the second line that $1-b \ge 0$ if μ is small enough.

- There exists a positive constant c_2 such that $h(x) \le -c_2 x\log(x)$ for $x \in (0, \sqrt{3}/2)$. Therefore, $h(a) = h((3/2)\sqrt{p}) \le -c_2(3/2)\sqrt{p}\log((3/2)\sqrt{p})$. As $(3/2)\sqrt{p} \ge p^\alpha$ for $0 < p \le 1/3$, we obtain $h(a) \le -c_2\alpha(3/2)\sqrt{p}\log(p) \le -c_3\log(p)$ for some absolute constant c_3. We therefore obtain $h(a)(1-b) \le 24c_3\mu p$.

- We have

$$\begin{aligned} c-a &= \frac{1-(3/2)\sqrt{p}}{1+2\mu p/\log(p)} - (1-(3/2)\sqrt{p}) \\ &= (1-(3/2)\sqrt{p})\frac{-2\mu p/\log(p)}{1+2\mu p/\log(p)} \le -4\mu p/\log(p), \end{aligned}$$

as $1 + 2\mu p/\log(p) \geq 1/2$. Also, $c \geq a \geq 1/8$ and $|h'(x)| \leq -c_0 \log(1-x)$ for $x \in (1/8, 1)$. Therefore, $|h(c) - h(a)| \leq -c_0 \log(1-c)|c-a| \leq 4\log(1-c)\mu p/\log(p)$. Finally, we have, if μ is small enough, using that $p \in (0, 1/3)$,

$$1 - c = \frac{(3/2)\sqrt{p} + 2\mu p/\log(p)}{1 + 2\mu p/\log(p)} \geq (3/8)\sqrt{p} \geq p^{\beta},$$

for some $\beta > 0$. Therefore, $|h(c) - h(a)| \leq c_4 \mu p$ for some absolute constant c_4.

As $0 < b \leq 8$, we finally obtain that there exists an absolute constant c_5 such that $F_\mu(p) \leq c_5 \mu p$ for $p \in (0, 1/3)$ and μ small enough. The conclusion is obtained by taking μ sufficiently small.

□

C.2 Proof of Theorem 4.6

Upperbound on $t_\lambda(B)$ By [ALS13, Lemma 5] for any $t \geq 0$, $h(B,t) \leq h(A, t+\gamma) + 2\gamma$. Therefore, according to Proposition 4.4, we have for $t^*(A) \leq t + \gamma < \tau(M)$,

$$h(t, B) \leq \frac{(t+\gamma)^2}{\tau(M)} + t^*(A)\left(1 + \frac{t^*(A)}{\tau(M)}\right) + 2\gamma.$$

Therefore, $h(t,B) < \lambda t$ if $\frac{(t+\gamma)^2}{\tau(M)} + t^*(A)\left(1 + \frac{t^*(A)}{\tau(M)}\right) + 2\gamma < \lambda t$. A straightforward computation shows that this is the case if $\gamma \leq t^*(A) \leq C_0 \lambda^2 \tau(M)$ for some absolute constant C_0 and $t_0 < t + \gamma < t_1$ with (using $\sqrt{1-u} \geq 1 - u$ for $u \in [0,1]$),

$$\begin{aligned} t_0 &:= \frac{\tau(M)\lambda}{2}\left(1 - \sqrt{1 - \frac{4}{\lambda^2 \tau(M)}\left(t^*(A)\left(1 + \frac{t^*(A)}{\tau(M)}\right) + (2+\lambda)\gamma\right)}\right) \\ &\leq \frac{2t^*(A)}{\lambda}\left(1 + \frac{t^*(A)}{\tau(M)}\right) + \frac{6\gamma}{\lambda} \end{aligned}$$

and $t_1 \geq \tau(M)\lambda/2$. Therefore, $t_\lambda(B) \leq \frac{2t^*(A)}{\lambda}\left(1 + \frac{t^*(A)}{\tau(M)}\right) + \frac{6\gamma}{\lambda}$, as long as $t_{\max} < \tau(M)\lambda/2 - \gamma$.

Lowerbound on $t_\lambda(A)$ in the noise-free case Assume that $\varepsilon(A)$ is sufficiently small so that Proposition 3.5 holds. Let $q \in M$ with $\varepsilon(A) = d(q, A)$. One has $q = \pi_M(x)$ for some $x \in \mathrm{Conv}_{t^*(A)}(A)$, so that

$$\begin{aligned} d(x, A) &\geq d(q, A) - \|x - q\| \geq \frac{t^*(A)}{\left(1 + C_0 \frac{\varepsilon(A)}{\tau(M)}\right)} - \frac{t^*(A)^2}{\tau(M)} \text{ by Proposition 3.5 and Lemma 3.3.} \\ &\geq t^*(A)\left(1 - C_0 \frac{\varepsilon(A)}{\tau(M)} - \frac{t^*(A)}{\tau(M)}\right) \\ &\geq t^*(A)\left(1 - C_0 \frac{2t^*(A)}{\tau(M)} - \frac{t^*(A)}{\tau(M)}\right) \geq t^*(A)\left(1 - C_1 \frac{t^*(A)}{\tau(M)}\right), \end{aligned}$$

where we used at the last line that $\varepsilon(A) \leq 2t^*(A)$ is $\varepsilon(A)/\tau(M)$ is sufficiently small by Proposition 3.5. As $x \in \mathrm{Conv}_{t^*(A)}(A)$, we have,

$$h(t^*(A), A) \geq t^*(A) \left(1 - C_1 \frac{t^*(A)}{\tau(M)}\right). \tag{C.16}$$

Therefore, if $\lambda \leq 1 - C_1 t^*(A)/\tau(M)$ and $t^*(A) < t_{\max}$, then $t_\lambda(A) \geq t^*(A)$.

Lowerbound on $t_\lambda(A)$ in the tubular noise case By [ALS13, Lemma 5] for any $t \geq \gamma$,

$$h(B, t) \geq h(A, t - \gamma) - 2\gamma. \tag{C.17}$$

Plugging $t = t^*(A) + \gamma$, and using (C.16),

$$h(B, t^*(A) + \gamma) \geq t^*(A) \left(1 - C_1 \frac{t^*(A)}{\tau(M)}\right) - 2\gamma. \tag{C.18}$$

This quantity is larger than $\lambda(t^*(A) + \gamma)$ as long as

$$C_1 \frac{t^*(A)}{\tau(M)} \leq 1 - \lambda - (2 + \lambda) \frac{\gamma}{t^*(A)}. \tag{C.19}$$

If $\gamma \leq (1 - \lambda)\frac{t^*(A)}{6}$ and $C_1 \frac{t^*(A)}{\tau(M)} \leq \frac{1-\lambda}{2}$, then (C.19) is satisfied, giving the desired lowerbound on $t_\lambda(B)$ under those two conditions, should $t^*(A) + \gamma$ be smaller than $t_{\max}$.

C.3 Proof of Corollaries 4.7, 4.9, 4.10

Lemma C.6. *Let $A \subset M$ be a finite set of cardinality n. Then,*

$$\varepsilon(A) \geq c_d \tau(M) n^{-1/d}. \tag{C.20}$$

Proof of Lemma C.6. As $M \subset \bigcup_{x \in A} \mathcal{B}(x, \varepsilon(A))$, one has $\mathrm{Vol}(M) \leq n c_d \varepsilon(A)^d$. Lemma III.24 and Proposition III.25 in [Aam17] imply that there exists a constant C_d such that $\mathrm{Vol}(M) \geq C_d \tau(M)^d$, thus leading to the conclusion. □

Proof of Corollary 4.7. By equation (B.9), if $t_\lambda(\mathbb{X}_n) \geq t^*(\mathbb{Y}_n) - \gamma$, then

$$d_H(\,\mathrm{Conv}_{t_\lambda(\mathbb{X}_n)}(\mathbb{X}_n), M) \leq \gamma + \frac{(t_\lambda(\mathbb{X}_n) + \gamma)^2}{\tau(M)}. \tag{C.21}$$

This relation holds (we even have $t_\lambda(\mathbb{X}_n) \geq t^*(\mathbb{Y}_n) + \gamma$) as long as Conditions 1, 2 and 3 of Theorem 4.6 are satisfied. If $\gamma < \eta\,(\log n/n)^{2/d}$ and $\tau_{\min} > 2t_{\max}/\lambda$, Conditions 1 and 2 are satisfied as long as $t^*(\mathbb{Y}_n)$ is small enough with respect to λ, $t_{\max}$ and $\tau(M)$ and n is large enough. Also, by

Lemma C.6 and Proposition 3.5, Condition 3 is satisfied as long as n is large enough. Therefore, Conditions 1, 2 and 3 are satisfied with probability $1-c_{d,\tau_{\min},f_{\min},\lambda,t_{\max}}\exp(-C_{d,\tau_{\min},f_{\min},\lambda,t_{\max}}n)$, according to Propositions 3.5 and 3.6. Therefore, (C.21) holds with high probability, and one obtains the conclusion by using the upper bound in Theorem 4.6, Proposition 3.5 and the fact that $\mathbb{E}[\varepsilon(\mathbb{Y}_n)^2]$ is of order $(\log n/n)^{2/d}$. □

Proof of Corollary 4.9. For the sake of simplicity, we prove the Corollary for $\eta = 0$ (no noise), but the extension to the noise case is made with similar ideas than in the previous proof. According to [CCSL09, Theorem 4.6], if $\varepsilon(\mathbb{X}_n) < \tau(M)/17$ and $4\varepsilon(\mathbb{X}_n) \leq t < \tau(M) - 3\varepsilon(\mathbb{X}_n)$, then $\mathcal{C}_t(\mathbb{X}_n) \simeq M$. Also, according to Theorem 4.6 and Proposition 3.5, if $\varepsilon(\mathbb{X}_n)$ is small enough with respect to λ, $t_{\max}$ and $\tau(M)$, then

$$5t_\lambda(\mathbb{X}_n) \geq 5t^*(\mathbb{X}_n) \geq 5\varepsilon(\mathbb{X}_n)\left(1 - C_0\frac{\varepsilon(\mathbb{X}_n)}{\tau(M)}\right) \geq 4\varepsilon(\mathbb{X}_n) \text{ and} \tag{C.22}$$

$$\begin{aligned} 5t_\lambda(\mathbb{X}_n) &\leq 10\frac{t^*(\mathbb{X}_n)}{\lambda}\left(1 + \frac{t^*(\mathbb{X}_n)}{\tau(M)}\right) \leq 20\frac{t^*(\mathbb{X}_n)}{\lambda} \\ &\leq \tau(M) - 3\varepsilon(\mathbb{X}_n). \end{aligned} \tag{C.23}$$

Therefore, if $\varepsilon(\mathbb{X}_n)$ is small enough, then $M \simeq \mathcal{C}_{5t_\lambda(\mathbb{X}_n)}(\mathbb{X}_n)$. We conclude by using Proposition 3.6. □

Proof of Corollary 4.10. According to [BSW09, Theorem 3.2], for $A \subset M$, if $t < \tau(M)/2$ and $t \geq 10\varepsilon(A)$, then

$$\angle(T_p(A,t), T_pM) \leq C_0\frac{t}{\tau(M)} \tag{C.24}$$

for some absolute constant C_0. According to Theorem 4.6 and Proposition 3.5, and arguing as in the two previous proofs, $11t_\lambda(\mathbb{X}_n) > 10\varepsilon(\mathbb{X}_n)$ and $11t_\lambda(\mathbb{X}_n) < \tau(M)/2$ as long as $\varepsilon(\mathbb{X}_n) < C_{\tau(M),\lambda,t_{\max}}$. Therefore,

$$\begin{aligned} \mathbb{E}\angle(T_pM, T_p(\mathbb{X}_n, 11t_\lambda(\mathbb{X}_n)) &\leq 11C_0\frac{\mathbb{E}t_\lambda(\mathbb{X}_n)}{\tau(M)} + \mathbb{P}(\varepsilon(\mathbb{X}_n) > C_{\tau(M),\lambda,t_{\max}}) \\ &\leq C_{d,\tau_{\min},f_{\min},\lambda,t_{\max}}\left(\log n/n\right)^{1/d}, \end{aligned}$$

by Theorem 4.6 and Proposition 3.6. □