Minimax adaptive estimation in manifold inference

Vincent Divol

▶ **To cite this version:**

Vincent Divol. Minimax adaptive estimation in manifold inference. 2020. hal-02440881v2

HAL Id: hal-02440881

https://inria.hal.science/hal-02440881v2

Preprint submitted on 8 Jun 2020 (v2), last revised 26 Oct 2021 (v3)

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

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

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. This selection procedure actually allows to recover the sample rate of the set of observations, and can therefore be used as an hyperparameter in other settings, such as tangent space estimation.

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 d-dimensional 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 d [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

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

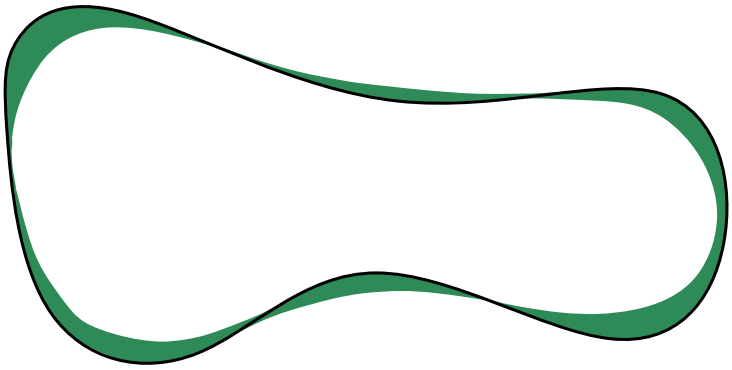


Figure 1 – The t-convex hull $\mathrm{Conv}_d(t;A)$ (in green) of a curve A (in black).

rates of convergence on a large class of models (see Section 2 for a precise definition). Our main contribution consists in introducing a manifold estimator $\hat{M}$ which is adaptive minimax (with respect to the Hausdorff distance d_H) on all the C^2-models with tubular noise introduced in [GPPVW12] and [AL18].

Our estimator is built by considering a family of estimators given by the t-convex hull $\mathrm{Conv}_d(t;\mathbb{X}_n)$ of the set of observations $\mathbb{X}_n$. For a given set $A \subset \mathbb{R}^D$, the (d-dimensional) t-convex hull $\mathrm{Conv}_d(t;A)$ is defined by

$$\mathrm{Conv}_d(t;A) := \bigcup_{\substack{\sigma \subset A,\ r(\sigma)\leq t \\ \dim(\sigma)\leq d}} \mathrm{Conv}(\sigma), \tag{1.1}$$

where $r(\sigma)$ is the *radius* of a set σ, i.e. the radius of the smallest enclosing ball of σ, $\dim(\sigma)$ is its dimension and $\mathrm{Conv}(\sigma)$ is its convex hull (see Definition 3.1). For $d = D$, 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 $d_H(\mathrm{Conv}_d(t;\mathbb{X}_n), M)$ of the t-convex hull $\mathrm{Conv}_d(t;\mathbb{X}_n)$ 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 $t_{\lambda,d}(\mathbb{X}_n)$ (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

$$h_d(t, A) = d_H(\,\mathrm{Conv}_d(t; A), A) \in [0, t] \text{ for } t \geq 0. \tag{1.2}$$

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 $\tau(M)$, $h_d(t, M) \leq t^2/\tau(M) \ll t$ for $t \ll \tau(M)$, 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 $h_d(t, \mathbb{X}_n)$ 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

$$t_{\lambda,d}(\mathbb{X}_n) := \sup\{t < t_{\max},\ h_d(t, \mathbb{X}_n) > \lambda t\}, \tag{1.3}$$

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

$$t^*(\mathbb{X}_n) \leq t_{\lambda,d}(\mathbb{X}_n) \leq \frac{2t^*(\mathbb{X}_n)}{\lambda}\left(1 + \frac{t^*(\mathbb{X}_n)}{\tau(M)}\right). \tag{1.4}$$

In particular, we are able to control the loss of $\mathrm{Conv}_d(t_{\lambda,d}(\mathbb{X}_n); \mathbb{X}_n)$ with high probability. By choosing $t_{\max}$ as a slowly decreasing function of n, and by using a preliminary estimator $\hat{d}$ of the dimension d, we obtain an estimator

$$\hat{M} := \mathrm{Conv}_{\hat{d}}(t_{\lambda,\hat{d}}(\mathbb{X}_n); \mathbb{X}_n)$$

which is adaptive on the whole collection of C^2-models as defined in Section 2 (see Corollary 4.7 and Remark 4.10 afterwards).

The estimator $\hat{M}$ is to our knowledge the first minimax adaptive, completely data-driven, manifold estimator. Our procedure actually allows us to estimate (up to a multiplicative constant) the sample rate $\varepsilon(\mathbb{X}_n)$. The parameter $t_{\lambda,\hat{d}}(\mathbb{X}_n)$ 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 tangent spaces of a manifold (see Corollary 4.9).

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 $t_{\lambda,d}(\mathbb{X}_n)$ is completely data-driven, computable in polynomial time (see Section 5), 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.

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 $\mathbb{X}_n = \mathrm{Conv}_d(0; \mathbb{X}_n)$) 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 setting of kernel density estimation.

Outline of the paper

The framework of minimax adaptive estimation as well as 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 $\mathrm{Conv}_d(t; \mathbb{X}_n)$ 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 $t_{\lambda,d}(\mathbb{X}_n)$, depending on two hyperparameters λ and $t_{\max}$, and we show the adaptivity of the estimator $\mathrm{Conv}_d(t_{\lambda,d}(\mathbb{X}_n); \mathbb{X}_n)$. We also discuss how the scale parameter $t_{\lambda,d}(\mathbb{X}_n)$ can be used as a scale parameter in the setting of tangent spaces estimation, leading to an adaptive procedure in this framework as well. We present some numerical illustrations of our procedure 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 $c_0, c_1, C_0, C_1, \dots$ will denote absolute constants in the following. If a constant depends on additional parameters $\alpha, \beta, \dots$, it will be denoted by $C_{\alpha,\beta,\dots}$.

Notations

The Euclidean norm in $\mathbb{R}^D$ is denoted by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ stands for the dot product. If $A \subset \mathbb{R}^D$ and $x \in \mathbb{R}^D$, then $d(x, A) := \inf_{y\in A} \|x - y\|$ is the distance to a set A while $\mathrm{diam}(A) := \sup_{x,y\in A} \|x - y\|$ is its diameter. Given $r \geq 0$, $\mathcal{B}(x, r)$ is the closed ball of radius r centered at x and we write $\mathcal{B}_A(x, r)$ for $\mathcal{B}(x, r) \cap A$. We let $\mathcal{C}^d$ be the set of C^2 compact connected d-dimensional submanifolds of $\mathbb{R}^D$ without boundary. If $M \in \mathcal{C}^d$ and $p \in M$, then 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 $d_H(A|B) := \sup_{x\in A} d(x, B)$. and the Hausdorff distance is then defined as $d_H(A, B) = \max\{d_H(A|B), d_H(B|A)\}$. The asymmetric Hausdorff distance verifies the following pseudo triangle inequality: for any sets $A, B, C \subset \mathbb{R}^D$, one has

$$d_H(A|C) \leq d_H(A|B) + d_H(B|C), \tag{2.1}$$

a fact we will use in the following. For $A \subset M$, we denote by $\varepsilon(A) := d_H(A, M)$ the sample rate of A.

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 C^2 compact submanifold without boundary has a positive reach $\tau(M) > 0$ (see [Fed59,

p. 432]). For $\tau_{\min} > 0$, we denote by $\mathcal{C}^d_{\tau_{\min}}$ the set of manifolds $M \in \mathcal{C}^d$ with reach larger than $\tau_{\min}$.

Minimax rates

Let $(\mathcal{Y}, \mathcal{H})$ be some measurable space and let $\mathcal{P}_0$ be a subset of the space of probability measures on $(\mathcal{Y}, \mathcal{H})$. Assume that there is a measurable function $\iota : (\mathcal{Y}, \mathcal{H}) \to (\mathcal{X}, \mathcal{G})$ such that we observe i.i.d. variables $X_1, \dots, X_n \sim \iota_\# P$ for some $P \in \mathcal{P}_0$. The tuple $(\mathcal{Y}, \mathcal{H}, \mathcal{P}_0, \mathcal{X}, \mathcal{G}, \iota)$ is a *statistical model*. Let θ be a functional of interest defined on $\mathcal{P}_0$, taking its value in some measurable space $(E, \mathcal{E})$ endowed with some measurable loss function $\rho : E \times E \to [0, \infty)$. We assume that $\mathcal{P}_0$ is written for $n \geq 0$ as an union $\bigcup_{q \in \mathcal{Q}} \mathcal{P}_{q,n}$, where the index $q \in \mathcal{Q}$ has to be thought as a measure of the regularity of the elements of $\mathcal{P}_0$. An estimator $\hat{\theta}$ on $\mathcal{P}_{q,n}$ is a measurable function $\mathcal{X}^{(\mathbb{N})} = \sqcup_{i \geq 1} \mathcal{X}^i \to E$, which may depend also on q. The risk of an estimator $\hat{\theta}$ on $\mathcal{P}_{q,n}$ given n observations is defined as

$$R_n(\hat{\theta}, \mathcal{P}_{q,n}) := \sup_{P \in \mathcal{P}_{q,n}} \mathbb{E}[\rho(\hat{\theta}(X_1, \dots, X_n), \theta(P))], \tag{2.2}$$

where $X_1, \dots, X_n$ are i.i.d. of law $\iota_\# P$. The minimax risk for the estimation of θ on $\mathcal{P}_{q,n}$ is then defined as

$$m_n(\theta; \mathcal{P}_{q,n}) := \inf_{\hat{\theta}} R_n(\hat{\theta}, \mathcal{P}_{q,n}), \tag{2.3}$$

where the infimum is taken over all estimators of θ on $\mathcal{P}_{q,n}$, i.e. the minimax risk is the best possible risk an estimator can attain uniformly on $\mathcal{P}_{q,n}$. An estimator $\hat{\theta}$ realizing the infimum $m_n(\theta; \mathcal{P}_{q,n})$ (up to a constant which does not depend on n) is called *minimax*. We say that an estimator $\hat{\theta}$ of θ on $\mathcal{P}_0$ (i.e. **not** depending on $q \in \mathcal{Q}$) is *minimax adaptive* on the whole collection $\mathcal{P}_0$ if

$$\sup_{q \in \mathcal{Q}} \limsup_{n \to \infty} \frac{R_n(\hat{\theta}, \mathcal{P}_{q,n})}{m_n(\theta; \mathcal{P}_{q,n})} < \infty. \tag{2.4}$$

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, with $f_{\max}$ possibly equal to $+\infty$. The set $\mathcal{P}^d_{\tau_{\min}, f_{\min}, f_{\max}}$ is the set of all distributions having for support a manifold $M \in \mathcal{C}^d_{\tau_{\min}}$, 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 $f_{\max}$. The statistical model is then build by letting $(\mathcal{X}, \mathcal{G}) = (\mathcal{Y}, \mathcal{H})$ be $\mathbb{R}^D$ endowed with its borelian σ-algebra and ι be the identity.*

Note that there are implicit constraints on the different parameters of the model. Indeed, by [NSW08, Proposition 6.1], the norm of the second fundamental form of a manifold M is

bounded by $1/\tau(M)$ and by [Alm86], this implies that the volume $\mathrm{Vol}(M)$ of M is larger than $\omega_d \tau(M)^d$, where ω_d is the volume of an unit d-sphere, with equality if and only if M is a d-dimensional sphere of radius $\tau(M)$. Hence, if P has a density f on M lowerbounded by $f_{\min}$, we have

$$1 = \int_M f(x)\mathrm{d}x \geq f_{\min} \mathrm{Vol}(M) \geq f_{\min}\omega_d \tau(M)^d,$$

with equality if and only if P is the uniform distribution on a d-sphere of radius $\tau(M)$. We therefore have the following lemma.

Lemma 2.2. *Let d be an integer smaller than D and $\tau_{\min}$, $f_{\min}$ be positive constants. Then, $\mathcal{P}^d_{\tau_{\min},f_{\min},+\infty}$ is empty for $f_{\min}\omega_d\tau^d_{\min} > 1$ and contains only uniform distributions on d-sphere of radius $\tau_{\min}$ if $f_{\min}\omega_d\tau^d_{\min} = 1$.*

A model containing only spheres is degenerate from a minimax perspective, as laws in the model are then characterized by only $d+1$ observations. To discard such a model, we will assume in the following that there exists a constant $\kappa < 1$ such that $f_{\min}\omega_d\tau^d_{\min} \leq \kappa^d$. Note that this is not restrictive as any $P \in \mathcal{P}^d_{\tau_{\min},f_{\min},+\infty}$ also belongs to $\mathcal{P}^d_{\tau'_{\min},f'_{\min},+\infty}$ for $\tau'_{\min} \leq \tau_{\min}$ and $f'_{\min} \leq f_{\min}$.

Definition 2.3 (Tubular noise model)**.** *Let d be an integer smaller than D and $\tau_{\min}$, $f_{\min}$, $f_{\max}$, γ be positive constants, with $f_{\max}$ possibly equal to $+\infty$. The set $\mathcal{P}^d_{\tau_{\min},f_{\min},f_{\max},\gamma}$ is the set of probability distributions P on $\mathbb{R}^D \times \mathbb{R}^D$ with first marginal P_1 in $\mathcal{P}^d_{\tau_{\min},f_{\min},f_{\max}}$ and second marginal P_2 supported on $\mathcal{B}(0,\gamma)$. The statistical model is then build by letting $(\mathcal{Y},\mathcal{H})$ be $\mathbb{R}^D \times \mathbb{R}^D$, $(\mathcal{X},\mathcal{G})$ be $\mathbb{R}^D$ (endowed with their borelian σ-algebras), and letting $\iota : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^D$ be the addition.*

Concretely, in the tubular noise model, we observe samples $X_1, \ldots, X_n$ of the form $X_i = Y_i + Z_i$, with the Y_is i.i.d. of law $P \in \mathcal{P}^d_{\tau_{\min},f_{\min},f_{\max}}$ and the Z_is are i.i.d. of norm smaller than γ, not necessarily independent from the Y_is. With a slight abuse of notation, we will identify $\mathcal{P}^d_{\tau_{\min},f_{\min},f_{\max}}$ with $\mathcal{P}^d_{\tau_{\min},f_{\min},f_{\max},0}$ in the following.

We let $\mathcal{P}^d(\kappa)$ be the union of the $\mathcal{P}^d_{\tau_{\min},f_{\min},f_{\max},\gamma}$ for $0 \leq d \leq D$ and $\tau_{\min}$, $f_{\min}$, $f_{\max}$, $\gamma > 0$ with $f_{\min}\omega_d\tau^d_{\min} \leq \kappa^d$. Also, let $\mathcal{P}(\kappa) := \bigcup_{d\leq D} \mathcal{P}^d(\kappa)$. For $P \in \mathcal{P}(\kappa)$, let $M(P)$ be equal to the support of its first marginal P_1. Then, M takes its values in 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 on this model have been studied in [AL18], following the works of [GPPVW12, KZ15]. We use the following parametrization of the set $\mathcal{P}^d(\kappa)$: let $\mathcal{Q}^d(\kappa)$ be the set of tuples $q = (\tau_{\min}, f_{\min}, f_{\max}, \eta)$, $\tau_{\min}, f_{\min}, f_{\max}, \eta > 0$ and $f_{\min}\omega_d\tau^d_{\min} \leq \kappa^d$. We let $\mathcal{P}^d_{q,n} = \mathcal{P}^d_{\tau_{\min},f_{\min},f_{\max},\gamma_n}$ for $\gamma_n = \eta(\log n/n)^{2/d}$.

Theorem 2.4. *Let $\kappa \in (0,1)$. For any $0 < d < D$ and $q = (\tau_{\min}, f_{\min}, f_{\max}, \eta) \in \mathcal{Q}^d(\kappa)$ with $f_{\max} < \infty$, we have for n large enough,*

$$\left(\frac{\eta}{2} + \frac{C(1-\kappa)}{(\omega_d f_{\min})^{2/d}\tau_{\min}}\right) \leq \liminf_n \frac{m_n(M; \mathcal{P}^d_{q,n})}{(\log n/n)^{2/d}} \leq \limsup_n \frac{m_n(M; \mathcal{P}_{q,n})}{(\log n/n)^{2/d}} \leq C_{q,d} \tag{2.5}$$

where C is an absolute constant and $C_{q,d}$ is a constant which depends on q and d.

The upper bound in the previous theorem is given by Theorem 2.9 in [AL18], whereas the constant in the lower bound follows from a careful adaptation of the proof of Theorem 1 in [KZ15], detailed in Appendix D.

Note that probability distributions in Theorem 2.4 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 techniques would be needed to obtain better estimators on such models. Note also that the t-convex hull estimators will be minimax on the model $\mathcal{P}^d_{\tau_{\min}, f_{\min}, +\infty, \gamma_n}$, for which the minimax rate is also equal to $(\log n/n)^{2/d}$ (the lower bound is clear, and the next section will show the upper bound).

As the parameter κ is fixed from now, we will drop the dependence in κ to ease the notations, i.e. $\mathcal{P} := \mathcal{P}(\kappa)$, $\mathcal{P}^d := \mathcal{P}^d(\kappa)$ and $\mathcal{Q}^d := \mathcal{Q}^d(\kappa)$.

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 $t \geq 0$. For $0 < d < D$, the d-dimensional t-convex hull of A is defined as*

$$\mathrm{Conv}_d(t; A) := \bigcup_{\substack{\sigma \subset A,\ r(\sigma) \leq t \\ \dim(\sigma) \leq d}} \mathrm{Conv}(\sigma), \tag{3.1}$$

where the dimension $\dim(\sigma)$ of a finite set σ is equal to its number of elements minus one.

In this section, we derive rates of convergence for $\mathrm{Conv}_d(t; \mathbb{X}_n)$, where $\mathbb{X}_n$ is a n-sample of law $P \in \mathcal{P}^d_{q,n}$.

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:

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

is measurable. Indeed, it can be written as

$$\bigcup_{I\subset\{1,\dots,n\}} \mathrm{Conv}(\{x_i\}_{i\in I}) \cap f_I(x_1,\dots,x_n)$$

where $f_I(x_1,\dots,x_n) = \emptyset$ if $r(\{x_i\}_{i\in I}) > t$ or $\dim(\{x_i\}_{i\in I}) > d$ and is equal to $\mathbb{R}^D$ otherwise. As the operations $\cup$, $\cap$, Conv are measurable [Aam17, Proposition III.7] and the function r is continuous [ALS13, Lemma 16], the measurability follows.

In order to obtain rates of convergence, we give a bound on the Hausdorff distance $d_H(\mathrm{Conv}_d(t;A), M)$ for a general subset $A \subset M$. First, [ALS13, Lemma 12] bounds 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 $y \in \mathrm{Conv}(\sigma)$. 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 $y \in \mathrm{Conv}(\sigma)$, 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 $d_H(\mathrm{Conv}_d(t;A)|M) \leq t^2/\tau(M)$ if $t < \tau(M)$, so that the set $\mathrm{Conv}_d(t;A)$ 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 $\mathrm{Conv}_d(t;A)$.

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

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

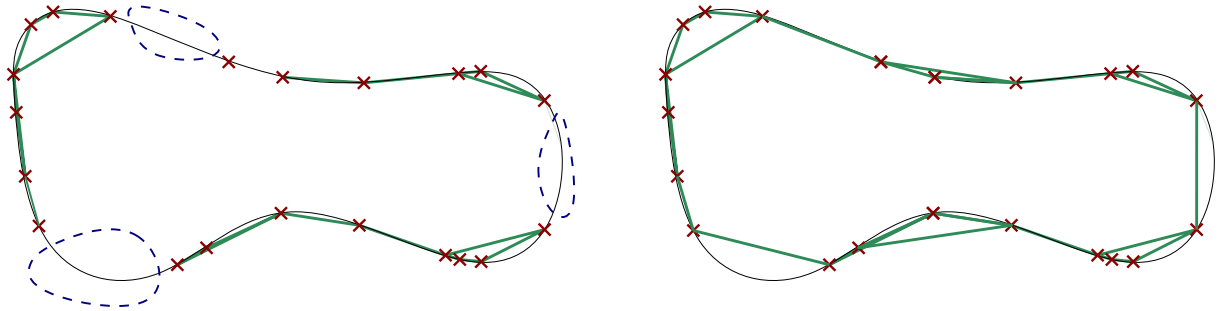
See Figure 2 for an illustration. Assume that $t^*(A) < +\infty$. Then, for $t^*(A) < t < \tau(M)$, and for any point $p \in M$, there exists $y \in \mathrm{Conv}_d(t;A)$ with $\pi_M(y) = p$. Therefore,

$$d(p, \mathrm{Conv}_d(t;A)) \leq \|y - p\| = d(y, M) \leq d_H(\mathrm{Conv}_d(t;A)|M).$$

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

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

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.



$t < t^*(A)$ $\qquad$ $t \geq t^*(A)$

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 $\varepsilon(A) \leq \tau(M)/78$, then*

$$\varepsilon(A)\left(1 - \frac{4}{3}\frac{\varepsilon(A)}{\tau(M)}\right) \leq t^*(A) \leq \varepsilon(A)\left(1 + 6\frac{\varepsilon(A)}{\tau(M)}\right). \tag{3.5}$$

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

Proposition 3.6. *Let $P \in \mathcal{P}_{d,\tau_{\min},f_{\min},+\infty}$ and let $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ be a n-sample of law P. If $r \leq \tau_{\min}/4$, then*

$$\mathbb{P}(\varepsilon(\mathbb{X}_n) > r) \leq \frac{8^d}{\alpha_d f_{\min} r^d} \exp(-n 2^d \alpha_d f_{\min} r^d), \tag{3.6}$$

where α_d is the volume of a unit d-ball. In particular, for n large enough

$$\mathbb{E}[\varepsilon(\mathbb{X}_n)^2] \leq 2\left(\frac{\log n}{\alpha_d f_{\min} n}\right)^{2/d} \leq 2\pi^2\left(\frac{\log n}{\omega_d f_{\min} n}\right)^{2/d}. \tag{3.7}$$

Proofs of Proposition 3.5 and Proposition 3.6 are found in Section B. 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, $n > 0$ and $q = (\tau_{\min}, f_{\min}, +\infty, \eta) \in \mathcal{Q}^d$. If $t_n = \left(\frac{3\log n}{2^d \alpha_d f_{\min} n}\right)^{1/d}$, then we have for n large enough*

$$R_n(\operatorname{Conv}_d(t_n; \mathbb{X}_n), \mathcal{P}^d_{q,n}) \leq \left(\frac{\log n}{n}\right)^{2/d}\left(\eta + \frac{4\pi^2}{\tau_{\min}(\omega_d f_{\min})^{2/d}}\right) \tag{3.8}$$

i.e. $\mathrm{Conv}_d(t_n; \mathbb{X}_n)$ *is a minimax estimator of* M *on* $\mathcal{P}^d_{q,n}$.

A proof of Theorem 3.7 is found in Section B.3.

4 Selection procedure for the t-convex hulls

Assuming that we have observed a n-sample $\mathbb{X}_n$ having a distribution $P \in \mathcal{P}^d_{q,n}$, 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 $d, \tau_{\min}, f_{\min}, \eta$, 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 $(\mathrm{Conv}_d(t; \mathbb{X}_n))_{t \geq 0}$. Assume first for the sake of simplicty that *the noise level* η *is null*. As the loss of the estimator $\mathrm{Conv}_d(t; \mathbb{X}_n)$ 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 $\mathrm{Conv}_d(t; \mathbb{X}_n)$ 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* d*-dimensional convexity defect function at scale* t *of* A *is defined as*

$$h_d(t, A) := d_H(\mathrm{Conv}_d(t; A), A). \tag{4.1}$$

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

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

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

Proof. As $h_d(t, A) \leq h_D(t, A)$, Point 1 follows from [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 $h_d(t, M)$ 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 $h_d(\cdot, \mathbb{X}_n)$ 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 $q = (\tau_{\min}, f_{\min}, f_{\max}, 0) \in \mathcal{Q}^d$ with $f_{\max} < +\infty$. Let $\mathbb{X}_n$ be a n-sample of law $P \in \mathcal{P}^d_{q,n}$. Fix $0 < \lambda < 1$. There exist positive constants t_0, C_0, C_1, C_2 depending on q and λ such that the following holds. Let, for $x > 0$, $\phi(x) = x^2 e^{-x}$ and $\psi(x) = \phi(x)/\log(1/\phi(x))$. Then, for n large enough and $0 < t \leq t_0$, we have*

$$h_d(t, \mathbb{X}_n) \geq \lambda t \text{ with probabilty larger than } 1 - C_0 \exp(-C_1 t^{-d} \psi(C_2 n t^d)). \tag{4.3}$$

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

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

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

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

$$\begin{aligned} h_d(t, A) &= d_H(\operatorname{Conv}_d(t; A), A) \\ &\leq d_H(\operatorname{Conv}_d(t; A), M) + d_H(M, \operatorname{Conv}_d(s; A)) + d_H(\operatorname{Conv}_d(s; A), A) \\ &\leq \frac{t^2}{\tau(M)} + \frac{s^2}{\tau(M)} + s. \end{aligned}$$

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 $h_d(t, \mathbb{X}_n)$ 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 $h_d(t, \mathbb{X}_n)$ is null with probability larger than, say, $1/2$, indicating that the lower bound in the previous range is close from 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*

$$t_{\lambda,d}(A) := \sup\{t < t_{\max},\ h_d(t, A) \geq \lambda t\}. \tag{4.5}$$

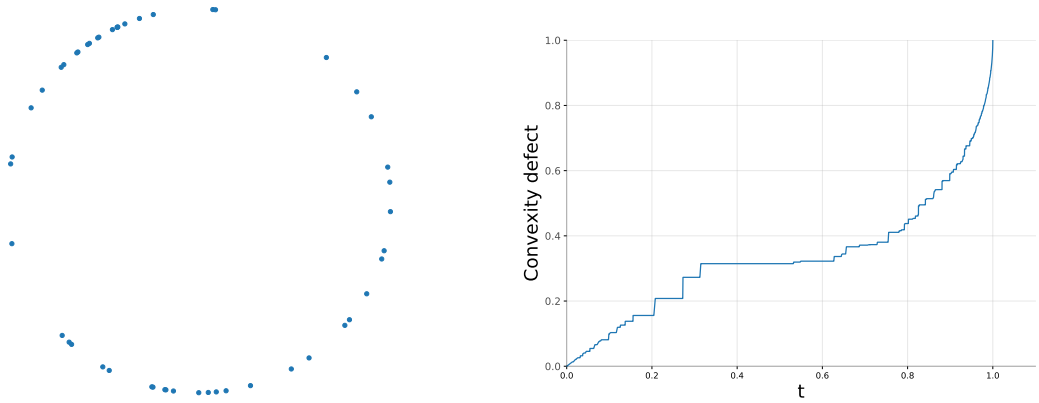


Figure 3 – The convexity defect function for 50 points $\mathbb{X}_n$ uniformly sampled on the unit circle. The two behaviors described by Proposition 4.3 and 4.4 are observed: $h_d(\cdot, \mathbb{X}_n)$ is linear at first, then a quadratic rate of growth appears. The threshold value is roughly at the value $t \simeq 0.3$, while $\varepsilon(\mathbb{X}_n)$ is equal to 0.32.

Propositions 4.3 and 4.4 prove that for any $\lambda < 1$, $t_{\lambda,d}(\mathbb{X}_n)$ 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 $M \in \mathcal{C}^d$. Let $A \subset M$ be a finite set with $\varepsilon(A) \leq \tau(M)/78$ and $B \subset \mathbb{R}^D$ with $d_H(A, B) \leq \gamma$. Assume that*

1. $t^*(A) + \gamma < t_{\max} < \frac{\lambda\tau(M)}{2}$,
2. $t^*(A) \leq c_\lambda \tau(M)$ *where* $c_\lambda = \min\left(\frac{13}{2}(1-\lambda), \frac{\lambda^2}{24}\right)$,
3. $\gamma \leq \frac{1}{6}(1-\lambda)t^*(A)$.

Then,

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

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

Corollary 4.7. *Let $0 < \lambda < 1$ and $t_{\max} = 1/\log(n)$. Let d be an integer smaller than D and $q = (\tau_{\min}, f_{\min}, +\infty, \eta) \in \mathcal{Q}^d$. Then, for n large enough*

$$R_n(\operatorname{Conv}_d(t_{\lambda,d}(\mathbb{X}_n); \mathbb{X}_n), \mathcal{P}^d_{q,n}) \leq \left(\frac{\log n}{n}\right)^{2/d} \left(\eta + \frac{121\pi^2}{\lambda^2(\omega_d f_{\min})^{2/d}\tau_{\min}}\right). \tag{4.7}$$

In particular, the estimator $\operatorname{Conv}_d(t_{\lambda,d}(\mathbb{X}_n); \mathbb{X}_n)$ is minimax adaptive on the scale of models given by $\mathcal{P}^d = \bigcup_{q\in\mathcal{Q}^d} \mathcal{P}^d_{q,n}$, i.e. we have

$$\sup_{q\in\mathcal{Q}^d} \limsup_n \frac{R_n(\operatorname{Conv}_d(t_{\lambda,d}(\mathbb{X}_n); \mathbb{X}_n), \mathcal{P}^d_{q,n})}{m_n(M, \mathcal{P}^d_{q,n})} \leq C_{\kappa,\lambda}. \tag{4.8}$$

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

Remark 4.8. Note that the previous result is of asymptotic nature. In particular, should n not be 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).

We now show that the parameter $t_{\lambda,d}(\mathbb{X}_n)$ 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]. For the sake of simplicity, we fix the level noise η at 0 in the next corollary, although it is possible to adapt the works of [BSW09] (and hence the next corollary) to samples with tubular noise. The angle between subspaces is denoted by $\angle$ (see Section A).

Corollary 4.9. *Let $0 < \lambda < 1$ and $t_{\max} = 1/\log(n)$. Let d be an integer smaller than D and $q = (\tau_{\min}, f_{\min}, +\infty, 0) \in \mathcal{Q}^d$. Then, for n large enough*

$$\sup_{P\in\mathcal{P}^d_{q,n}} \sup_{p\in M} \mathbb{E}\angle(T_pM, T_p(\mathbb{X}_n, 11t_{\lambda,d}(\mathbb{X}_n))) \leq \left(\frac{\log n}{n}\right)^{1/d} \frac{137\sqrt{2}\pi}{\lambda(\omega_d f_{\min})^{1/d}\tau_{\min}}, \tag{4.9}$$

where M denotes the underlying manifold of P.

This rate is the minimax rate (up to logarithmic factors) according to [AL19, Theorem 3]. A proof of Corollary 4.9 is found in Section C.3.

Remark 4.10. An issue with the estimators of the previous corollaries is that they still require the a priori knowledge of the dimension d of the manifold M. As a consequence, the estimators are only adaptive on $\mathcal{P}^d$, and not $\mathcal{P} = \bigcup_{0<d<D} \mathcal{P}^d$. To obtain adaptive estimators on $\mathcal{P}$, it is possible to use a parameter-free estimator $\hat{d}$ of d as a preliminary estimator, and then to use $t_{\lambda,\hat{d}}(\mathbb{X}_n)$ as a selected parameter. Such a dimension estimator is described in [BH19, Definition 5], and it satisfies $\mathbb{P}(\hat{d} \neq d) \leq 4\exp(2n^{1-(d+1)/(D+1)})$ for n large enough. This superpolynomial rate of convergence ensures that the manifold estimator $\operatorname{Conv}_{\hat{d}}(t_{\lambda,\hat{d}}(\mathbb{X}_n); \mathbb{X}_n)$ is adaptive on $\mathcal{P}$.

5 Numerical considerations

The selection procedure described in Section 4 amounts to compute the convexity defect function of the set $\mathbb{X}_n \subset \mathbb{R}^D$. To do so, we need for each simplex $\sigma \subset \mathbb{X}_n$ of dimension less than d to (i) compute its radius $r(\sigma)$ and (ii) compute $d_H(\mathrm{Conv}(\sigma)|\mathbb{X}_n)$. As the dimension of σ is less than d, its radius can be computed in constant time while [ABG+03] proposes an algorithm to compute the distance between a d-simplex and a family of m points with time complexity $O(Dm^{d+2})$. As there are $O(n^{d+1})$ simplexes of dimension less than d in $\mathbb{X}_n$, a direct use of this algorithm yields that $h_d(\cdot, \mathbb{X}_n)$ can be computed in $O(Dn^{2d+3})$. This complexity may be prohibitive for large n, but can be reduced by computing an approximation of the convexity defect function: the Hausdorff distance $d_H(\mathrm{Conv}(\sigma)|\mathbb{X}_n)$ can be approximated by computing the Hausdorff distance between a discretization of size $\eta r(\sigma)$ of $\mathrm{Conv}(\sigma)$ and $\mathbb{X}_n$ for some $\eta \in (0, 1)$, which can be done (naively) in $O(Dn\eta^{-d})$ time. Hence, the worst time complexity of the algorithm becomes $O(Dn^{d+2}\eta^{-d})$. We now argue that it suffices to compute the convexity defect function for $d = 1$ to select a good scale t, whatever the dimension d of the underlying manifold M is actually equal to.

Lemma 5.1. *Let $1 \leq d \leq D$ be an integer and let $c_d = \sqrt{\frac{1}{2} - \frac{1}{2d}}$. Let $B \subset \mathbb{R}^D$ and $t_{\max} > 0$, $c_d < \lambda < 1$. Then,*

$$t_{\lambda,d}(B) \leq t_{\lambda,1}(B) \leq t_{\lambda - c_d,d}(B). \tag{5.1}$$

Proof. A direct computation shows that if σ is a d-simplex of radius smaller than t, then the Hausdorff distance between $\mathrm{Conv}(\sigma)$ and the 1-skeleton of σ (the union of its edges) is bounded by $c_d t$. Hence, $h_d(t, B) \leq h_1(t, B) \leq h_d(t, B) + c_d t$. The conclusion follows from the definition of $t_{\lambda,d}(B)$. □

Hence, if some sets $A, B \subset M$ satisfy the conditions of Theorem 4.6 for λ and $\lambda - c_d$, then $t_{\lambda,1}(B)$ satisfies

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

and $\mathrm{Conv}_d(t_{\lambda,1}(\mathbb{X}_n); \mathbb{X}_n)$ is also a minimax adaptive estimator, while the time complexity for the computation of an η-approximation of $t_{\lambda,1}(\mathbb{X}_n)$ is $O(Dn^3\eta^{-1})$. If a cubic complexity is still too expensive, it is possible to only compute $d_H(\mathrm{Conv}(\sigma)|\mathbb{X}_n)$ for a random subset of L pairs σ in $\mathbb{X}_n$. The time complexity is then of order $O(DnL\eta^{-1})$, and the output of the algorithm is a function smaller than $h_1(\cdot, \mathbb{X}_n)$, so that the selected t will be larger than $t_{\lambda,1}(\mathbb{X}_n)$. If we have no guarantees on the output of this last algorithm, it appears in our experiments that it is similar to $h_1(\cdot, \mathbb{X}_n)$ for L significantly smaller than n^2.

As a numerical illustration of our procedure, we compute the convexity defect function $h_1(\cdot, \mathbb{X}_n)$ of three synthetic datasets: (a) $n_a = 10^3$ points uniformly sampled on the unit circle,

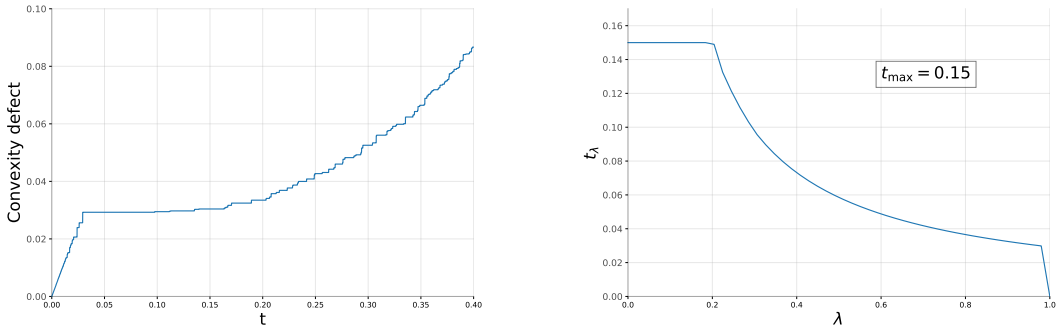
(b) $n_b = 10^4$ points sampled on a torus of inner radius 4 and outer radius 1, and (c) $n_c = 10^5$ points sampled on a swiss roll using the SciPy Python library [VGO+20] (which was also used to compute the Hausdorff distance between point clouds). The convexity defect functions (a), (b) and (c) were approximated using the algorithm described in the previous paragraph with parameters $\eta = 0.1$ and respectively $L_a = \infty$ (all pairs computed), $L_b = 10^6$ and $L_c = 10^7$. On each function, displayed in Figure 4, the behavior described in Section 4 is observed: first a linear growth up to a certain value, then a quadratic growth until the reach of the manifold (equal to 1 in the first two illustrations, and slightly larger than 3 for the swiss roll dataset). We then fix $t_{\max} = 0.5\,\mathrm{diam}(\mathbb{X}_n)/\log(n)$ and compute $t_{\lambda,1}(\mathbb{X}_n)$ for different values of λ. When λ is very close to 1, $t_{\lambda,1}(\mathbb{X}_n)$ is always 0, whereas it slowly increases as λ decreases, until reaching $t_{\max}$ at some value $\lambda_{\min}$. As a rule of thumb, we choose $\lambda_* = \frac{1+\lambda_{\min}}{2}$ and select the parameter $t_{\lambda_*,1}(\mathbb{X}_n)$, which is equal to $t_a = 0.049$, $t_b = 0.31$ and $t_c = 0.48$ in the different experiments (a), (b) and (c), while the sample rates $\varepsilon(\mathbb{X}_n)$ where evaluated (by oversampling) at $\varepsilon_a = 0.021$, $\varepsilon_b = 0.31$ and $\varepsilon_c = 0.33$.

6 Discussion and further works

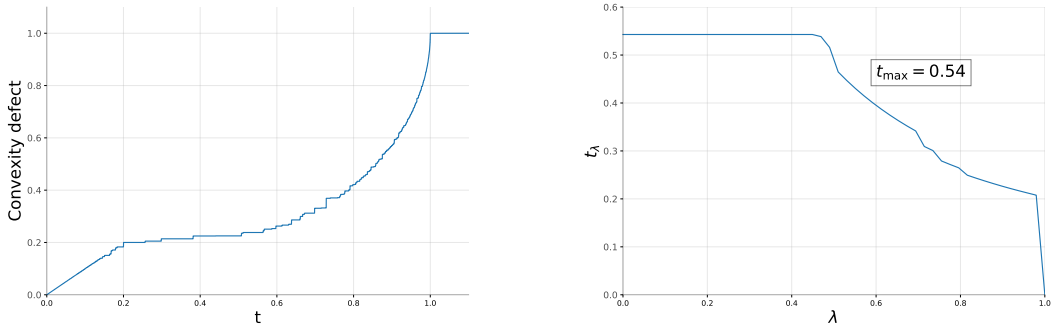
In this article, we introduced a particularly simple manifold estimator, based on an unique rule: add the convex hull of any subset of the set of observations which is of radius smaller than t. After proving that this leads to a minimax estimator for some choice of t, we explained how to select the parameter t by computing the convexity defect function of the set of observations. Surprisingly enough, the selection procedure allows to find a parameter $t_{\lambda,d}(\mathbb{X}_n)$ which is with high probability between, say, $\frac{1}{3}\varepsilon(\mathbb{X}_n)$ and $3\varepsilon(\mathbb{X}_n)$ (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 an adaptive tangent space estimator can be created thanks to $t_{\lambda,d}(\mathbb{X}_n)$. The main limitation to our procedure is its non-robustness to outliers. Indeed, even in the presence of one outlier in $\mathbb{X}_n$, the loss function $t \mapsto d_H(\,\mathrm{Conv}_d(t;\mathbb{X}_n), M)$ would be constant, equal to the distance between the outlier and the manifold M: with respect to the Hausdorff distance, all the estimators $\mathrm{Conv}_d(t;\mathbb{X}_n)$ 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 $\mathrm{Conv}_d(t;\mathbb{X}_n)$ instead of the Hausdorff distance.

Acknowledgements

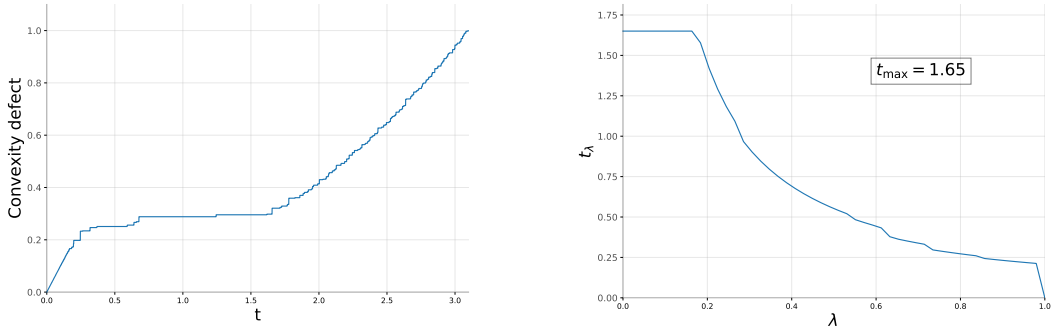
I am grateful to Fréderic Chazal (Inria Saclay) and Pascal Massart (Université Paris-Sud) for helpful discussions and valuable comments on both mathematical and computational aspects of this work. I would also like to thank Théo Lacombe for his thorough re-reading of the paper.



(a) 10^3 points on a circle



(b) 10^4 points on a torus



(c) 10^5 points on a swiss roll

Figure 4 – The convexity defect function of the datasets (a), (b) and (c), and the corresponding choices of $t_{\lambda,1}(\mathbb{X}_n)$ with respect to λ.

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.

[ABG+03] Helmut Alt, Peter Braß, Michael Godau, Christian Knauer, and Carola Wenk. Computing the hausdorff distance of geometric patterns and shapes. In *Discrete and computational geometry*, pages 65–76. Springer, 2003.

[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.

[Alm86] Fred Almgren. Optimal isoperimetric inequalities. *Indiana University mathematics journal*, 35(3):451–547, 1986.

[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.

[BC+09] Jonathan M Borwein, O Chan, et al. Uniform bounds for the complementary incomplete gamma function. *Mathematical Inequalities and Applications*, 12:115–121, 2009.

[BH19] Clément Berenfeld and Marc Hoffmann. Density estimation on an unknown submanifold. *arXiv preprint arXiv:1910.08477*, 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.

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

[Gal68] Robert G Gallager. *Information theory and reliable communication*, volume 2. Springer, 1968.

[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 43-6, pages 763–774, 2007.

[VGO+20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 2020.

A Properties of manifolds with reach constraints

In this section, M is a manifold in $\mathcal{C}^d$. 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 the distance $\|\pi_U - \pi_V\|_{\text{op}}$ 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 $\|p-q\| \le \tau(M)/2$. Then,*

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

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 map $\tilde{\pi}_p$ 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, we have $3\|p-q\|/\tau(M) > 1$, 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 diffeormorphism 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 $\tilde{\sigma} = \tilde{\pi}_p(\sigma)$. Assume that $p \in \mathrm{Conv}(\tilde{\sigma})$. Then,*

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

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)

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

where we used that $\|\tilde{x}_i - p\| \leq 2r(\tilde{\sigma})$ as $p \in \mathrm{Conv}(\tilde{\sigma})$. Let $\tilde{z}$ be the center of the minimum enclosing ball of $\tilde{\sigma}$. Write $\tilde{z} = \sum_{j=0}^k \lambda_j \tilde{x}_j$ as a convex combination of the points of $\tilde{\sigma}$ 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 $M \in \mathcal{C}^d$ be a d-dimensional manifold and let $A \subset M$ be a finite set. If $t^*(A) \leq \tau(M)/36$, then*

$$t^*(A) \leq \varepsilon(A)\left(1 + 6\frac{\varepsilon(A)}{\tau(M)}\right) \quad \textit{and} \tag{B.1}$$

$$t^*(A) \geq \varepsilon(A)\left(1 - \frac{4}{3}\frac{\varepsilon(A)}{\tau(M)}\right). \tag{B.2}$$

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

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

We have $t^*(A) = \sup_{p\in M} t^*(p, A) \leq \tau(M)/36$. Let $p \in M$ be such that $t^*(p, A) = t^*(A)$. Let $\sigma(p)$ be a simplex of A (of dimension less than d) 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. *Under the assumption $t^*(A) = t^*(p, A) \leq \tau(M)/36$, the set $\tilde{A}_p$ does not lie on a hyperplane of $\tilde{T}_pM$.*

Proof. Assume that $\tilde{A}_p$ lies on some hyperplane H of $\tilde{T}_pM$. Then, as $t^*(p, A) \leq \tau(M)/36$, we have $p \in \operatorname{Conv}(\tilde{\sigma}_p) \subset \operatorname{Conv}(\tilde{A}_p) \subset H$ by assumption. Therefore, the hyperplane H contains p. Consider a point $\tilde{q} \in \tilde{T}_pM$ 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 $r(\sigma(p))$. Therefore, if $t^*(A) = r(\sigma(p)) \leq \tau(M)/36$, then by Lemma A.5,

$$\begin{aligned}
\sigma(q) &\subset \mathcal{B}_M\left(q, 2r(\sigma(q))\left(1 + \frac{r(\sigma(q))}{2\tau(M)}\right)\right) \\
&\subset \mathcal{B}_M\left(q, \frac{\tau(M)}{18}\left(1 + \frac{1}{72}\right)\right) \subset \mathcal{B}_M\left(q, \frac{\tau(M)}{8}\right) \\
&\subset \mathcal{B}_M\left(p, \|p - q\| + \frac{\tau(M)}{8}\right) \subset \mathcal{B}_M\left(p, \frac{\tau(M)}{4}\right)
\end{aligned}$$

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, $\tilde{\pi}_p(y) \in \operatorname{Conv}(\tilde{A}_p) \subset H$. Therefore, recalling that $\tilde{q} - p$ is orthogonal to H, we have

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

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. □

Hence, there exists a Delaunay triangulation of $\tilde{A}_p$, which we will consider in the following. If $t^*(p, A) \leq \tau(M)/36$, 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 $p \in \tilde{\pi}_p(\operatorname{Conv}(\sigma(p))) \subset \operatorname{Conv}(\tilde{A}_p)$ and that there exists a d-simplex $\tilde{\sigma}_0$ in a Delaunay triangulation of $\tilde{A}_p$ with $p \in \operatorname{Conv}(\tilde{\sigma}_0)$. We denote by σ_0 be the corresponding d-simplex in A.

Lemma B.3. *Assume that $p \in M$ satisfies $t^*(p, A) \leq \tau(M)/36$ and that there exists $\tilde{y} \in \tilde{T}_p M$ 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 and Lemma 3.3. 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 that $t^*(A) = t^*(p, A) \leq \varepsilon(A)\left(1 + 6\frac{\varepsilon(A)}{\tau(M)}\right)$ in both cases. □

Proof of Lemma B.3. According to Point 2 in Lemma A.3, if we have $t^*(p, A) \leq \tau(M)/36$, then there exists $y \in \mathcal{B}_M(p, \tau(M)/4)$ with $\tilde{\pi}_p(y) = \tilde{y}$. As the projection is 1-Lispchitz, we have $d(\tilde{y}, \tilde{A}_p) \leq d(y, A \cap \mathcal{B}(p, \tau(M)/4))$. 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 of dimension less than d with $r(\sigma(p)) = t^*(p, A)$ and such that $p \in \pi_M(\mathrm{Conv}(\sigma(p)))$. By Lemma A.5, $\sigma(p) \subset \mathcal{B}(p, \tau(M)/4)$. Therefore, for $x \in \sigma(p)$, we have

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

From 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 exists a simplex $\sigma(p)$ of dimension less than d with $r(\sigma(p)) \leq t^*(A)$ and $x \in \mathrm{Conv}(\sigma(p))$ with $\pi_M(x) = p$. 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 $t^*(A) \leq \tau(M)/36$,

$$\begin{aligned}
\varepsilon(A) &\leq t^*(A)\left(1 + \frac{\varepsilon(A)\left(1 + 6\frac{\varepsilon(A)}{\tau(M)}\right)}{\tau(M)}\right) \\
&\leq t^*(A)\left(1 + \frac{4}{3}\frac{\varepsilon(A)}{\tau(M)}\right),
\end{aligned}$$

so that $\varepsilon(A)\left(1 - \frac{4}{3}\frac{\varepsilon(A)}{\tau(M)}\right) \leq t^*(A)$ 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 $\tilde{y} \in \mathcal{B}_{\tilde{T}_p M}(p, \varepsilon(A))$. 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,

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

By Lemma B.4, this implies that

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

By Carathéodory's theorem, there exists a d-simplex $\tilde{\sigma}_p \subset \tilde{\pi}_p(A \cap \mathcal{B}(p, 15\varepsilon(A)/7))$ such that $p \in \mathrm{Conv}(\tilde{\sigma}_p)$. Let σ_p be the corresponding simplex in $A \cap \mathcal{B}(p, 15\varepsilon(A)/7)$. If $15\varepsilon(A)/7 < \tau(M)$, then there is $x \in \mathrm{Conv}(\sigma_p)$ 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

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

as long as $\varepsilon(A) < 7\tau(M)/24$, thus showing (B.5) with $c_0 = 15/7$ and $c_1 = 7/24$. If $\varepsilon(A) \leq \tau(M)/78$, then $t^*(A) \leq c_0 \tau(M)/78 \leq \tau(M)/36$, concluding the proof of Proposition 3.5. □

B.2 Proof of Proposition 3.6

Proof of Equation (3.6) is found in [Aam17, Lemma III.23]. To obtain Equation (3.7), we use that for $L > 0$,

$$\begin{aligned}
\mathbb{E}[\varepsilon(\mathbb{X}_n)^2] &= \mathbb{E}[\varepsilon(\mathbb{X}_n)^2 \mathbf{1}\{\varepsilon(\mathbb{X}_n) \leq L\}] + \mathbb{E}[\varepsilon(\mathbb{X}_n)^2 \mathbf{1}\{L \leq \varepsilon(\mathbb{X}_n) \leq \tau_{\min}/4\}] \\
&\quad + \mathbb{E}[\varepsilon(\mathbb{X}_n)^2 \mathbf{1}\{\tau_{\min}/4 \leq \varepsilon(\mathbb{X}_n)\}] \\
&\leq L^2 + \int_{L^2}^{(\tau_{\min}/4)^2} \mathbb{P}(\varepsilon(\mathbb{X}_n)^2 > u) \mathrm{d}u + \operatorname{diam}(M)^2 \mathbb{P}(\varepsilon(\mathbb{X}_n) > \tau_{\min}/4) \\
&\leq L^2 + \int_{L^2}^{+\infty} \frac{8^d}{\alpha_d f_{\min} u^{d/2}} \exp(-n 2^d \alpha_d f_{\min} u^{d/2}) \mathrm{d}u \\
&\quad + \operatorname{diam}(M)^2 \frac{16^d}{\alpha_d f_{\min} \tau_{\min}^{d/2}} \exp(-n 4^d \alpha_d f_{\min} \tau_{\min}^{d/2}).
\end{aligned}$$

As $\operatorname{diam}(M)$ is bounded by a constant depending on $d, f_{\min}, \tau_{\min}$ (see [AL18, Lemma 2]), the last term is negligible in front of $(\log n/n)^{2/d}$ if n is large enough with respect to the parameters of the model. Also, by a change of variables, the second term is equal to

$$\begin{aligned}
&\int_{n 2^d \alpha_d f_{\min} L^d}^{+\infty} \frac{16^d n \alpha_d f_{\min}}{\alpha_d f_{\min} v} \exp(-v) \frac{1}{(n 2^d \alpha_d f_{\min})^{2/d}} \frac{2}{d} v^{2/d-1} \mathrm{d}v \\
&\quad = \frac{2}{d} \frac{16^d}{(2^d \alpha_d f_{\min})^{2/d}} n^{1-2/d} \int_{n 2^d \alpha_d f_{\min} L^d}^{+\infty} v^{2/d-2} \exp(-v) \mathrm{d}v \\
&\quad \leq \frac{1}{2d} \frac{16^d}{(\alpha_d f_{\min})^{2/d}} n^{1-2/d} (n 2^d \alpha_d f_{\min} L^d)^{2/d-1} \exp(-n 2^d \alpha_d f_{\min} L^d) \\
&\quad \leq 2 \frac{8^d}{d \alpha_d f_{\min}} L^{2-d} \exp(-n 2^d \alpha_d f_{\min} L^d),
\end{aligned}$$

where, at the second to last line, we used a classical bound on the incomplete Gamma function (see [BC+09, Theorem 2.1]). Letting $L^d = a \log n/(n 2^d \alpha_d f_{\min})$ for $a > 0$, we obtain

$$\begin{aligned}
\mathbb{E}[\varepsilon(\mathbb{X}_n)^2 \mathbf{1}\{\varepsilon(\mathbb{X}_n) \leq \tau_{\min}/4\}] &\leq L^2 \left(1 + 2 \frac{8^d}{d \alpha_d f_{\min}} \frac{L^{-d}}{n^a}\right) \\
&\leq \left(\frac{a \log n}{n 2^d \alpha_d f_{\min}}\right)^{2/d} \left(1 + \frac{16^d}{d} \frac{2}{a n^{a-1} \log n}\right).
\end{aligned}$$

Choosing $1 \leq a < 2$ yields that for n large enough,

$$\begin{aligned}
\mathbb{E}[\varepsilon(\mathbb{X}_n)^2] &\leq \left(\frac{\log n}{n \alpha_d f_{\min}}\right)^{2/d} + \operatorname{diam}(M)^2 \frac{16^d}{\alpha_d f_{\min} \tau_{\min}^{d/2}} \exp(-n 4^d \alpha_d f_{\min} \tau_{\min}^{d/2}) \\
&\leq 2 \left(\frac{\log n}{n \alpha_d f_{\min}}\right)^{2/d}.
\end{aligned}$$

Also, note that $(\omega_d/\alpha_d)^{2/d} \leq \pi^2$, yielding the second inequality in (3.7).

B.3 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,*

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

Proof. Let $\sigma \subset B$ be a simplex of dimension less than d with $r(\sigma) \leq t$. For each $y \in \sigma$, let $x \in A$ with $\|x - y\| \leq \gamma$. By doing so, we create a non-empty simplex $\xi \subset A$ of dimension less than d with $d_H(\sigma|\xi) \leq \gamma$. 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 $t^*(A) < t - \gamma < t + \gamma < \tau(M)$, using (2.1), Lemma B.5 and (3.4),

$$\begin{aligned}
d_H(\operatorname{Conv}_d(t; B)|M) &\leq d_H(\operatorname{Conv}_d(t; B) | \operatorname{Conv}_d(t + \gamma; A)) + d_H(\operatorname{Conv}_d(t + \gamma; A)|M) \\
&\leq \gamma + \frac{(t + \gamma)^2}{\tau(M)} \text{ and} \\
d_H(M | \operatorname{Conv}_d(t; B)) &\leq d_H(M | \operatorname{Conv}_d(t - \gamma; A)) + d_H(\operatorname{Conv}_d(t - \gamma; A) | \operatorname{Conv}_d(t; B)) \\
&\leq \frac{(t - \gamma)^2}{\tau(M)} + \gamma,
\end{aligned}$$

so that

$$d_H(\operatorname{Conv}_d(t; B), M) \leq \gamma + \frac{(t + \gamma)^2}{\tau(M)}. \tag{B.9}$$

Let $q = (\tau_{\min}, f_{\min}, +\infty, \eta) \in \mathcal{Q}^d$, let $P \in \mathcal{P}^d_{q,n}$ with underlying manifold M and let $\mathbb{X}_n$ be a n-sample of law $\iota_\# P$, with $\mathbb{Y}_n$ the corresponding sample of law P_1, the first marginal of P. Then, for $0 \leq t < \tau(M) - \gamma$,

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

By Proposition 3.5, if $\varepsilon(\mathbb{Y}_n) \leq \tau(M)/78$, then $t^*(\mathbb{Y}_n) \geq t$ implies that

$$\varepsilon(\mathbb{Y}_n) \geq t \left(1 + 6 \frac{\varepsilon(\mathbb{Y}_n)}{\tau(M)}\right)^{-1} \geq \frac{13}{14} t.$$

Therefore, if $t \leq \tau(M)/78$ and $t^*(\mathbb{Y}_n) \geq t$ then $\varepsilon(\mathbb{Y}_n) \geq \frac{13}{14}t$. 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 $t \leq \tau(M)/78$,

$$\mathbb{E}d_H(\operatorname{Conv}_d(t;\mathbb{X}_n), M) \leq \gamma + \frac{(t+\gamma)^2}{\tau(M)} + C_{d,\tau_{\min},f_{\min}} \frac{\exp(-2^d\alpha_d f_{\min} n(t-\gamma)^d)}{(t-\gamma)^d}. \tag{B.10}$$

In particular, by letting $t = \left(\frac{3\log n}{2^d\alpha_d f_{\min} n}\right)^{1/d}$, if $\gamma \leq \eta\,(\log n/n)^{2/d}$, we obtain

$$\begin{aligned}
&\mathbb{E}d_H(\operatorname{Conv}_d(t;\mathbb{X}_n), M) \\
&\leq \left(\frac{\log n}{n}\right)^{2/d}\left(\eta + \frac{1}{\tau_{\min}}\left(\left(\frac{3}{2^d\alpha_d f_{\min}}\right)^{1/d} + \eta\left(\frac{\log n}{n}\right)^{1/d}\right)^2\right) + C'_{d,\tau_{\min},f_{\min}}\frac{n^{-2}}{\log n} \\
&\leq \left(\frac{\log n}{n}\right)^{2/d}\left(\eta + \frac{1}{\tau_{\min}}\left(\frac{4}{2^d\alpha_d f_{\min}}\right)^{2/d}\right) \text{ if } n \text{ is large enough} \\
&\leq \left(\frac{\log n}{n}\right)^{2/d}\left(\eta + \frac{1}{\tau_{\min}}\pi^2\frac{4^{2/d}}{4}\left(\frac{1}{\omega_d f_{\min}}\right)^{2/d}\right) \text{ as } (\omega_d/\alpha_d)^{2/d} \leq \pi^2 \\
&\leq \left(\frac{\log n}{n}\right)^{2/d}\left(\eta + \frac{4\pi^2}{\tau_{\min}}\left(\frac{1}{\omega_d f_{\min}}\right)^{2/d}\right).
\end{aligned}$$

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

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

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 $k = 1, \ldots, K$, 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 $h_d(t, \mathbb{X}_n) < \lambda t$, then E_k is satisfied for $k = 1, \ldots, K$.*

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

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

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

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

Proof. Let Y_1, Y_2 be two independent random variables sampled according to P, conditioned on

being in B'_k. Then,

$$\begin{aligned}
\mathbb{P}(F_k|N_k=2) &= \mathbb{P}(N_{B'_k}<2|N_k=2) \\
&\qquad + \mathbb{P}(N_{B'_k}=2 \text{ and } r(\mathbb{X}_n\cap U_k)<\lambda t|N_k=2) \\
&= 1-\mathbb{P}(N_{B'_k}=2 \text{ and } r(\mathbb{X}_n\cap U_k)\geq\lambda t|N_k=2) \\
&= 1-\mathbb{P}(N_{B'_k}=2|N_k=2)\mathbb{P}(r(\mathbb{X}_n\cap B'_k)\geq\lambda t|N_{B'_k}=2) \\
&= 1-\left(\frac{P(B'_k)}{P(U_k)}\right)^2\mathbb{P}(r(\{Y_1,Y_2\})\geq\lambda t) \\
&\leq 1-\left(KC_df_{\min}t^d\right)^2\mathbb{P}(\|Y_1-Y_2\|\geq 2\lambda t) \\
&\leq 1-C_1\mathbb{P}(\|Y_1-Y_2\|\geq 2\lambda t),
\end{aligned}$$

where we used [Aam17, Lemma III.23] at the second to last line and Lemma C.1 at the last line. Let x_1, x_2 be two antipodal points on B'_k. If $\|x_i-Y_i\|\leq(1-\lambda)t$ for $i=1,2$, then $\|Y_1-Y_2\|\geq 2\lambda t$. 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,

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

where we used [Aam17, Lemma III.23]. This concludes the proof. □

We finally obtain

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

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 $n \geq 2$,

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

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

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

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,

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

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

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

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

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

Therefore,

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

Note that, should condition (C.11) be not satisfied, then the right-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 right-hand side of (C.13) is larger than 1 in this case. As the left-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 $p\ge (n/2K)^2e^{-2n/K}\ge (C_8nt^d)^2\exp(-C_8nt^d)$ for some constant C_8. 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 [Gal68, Exercise 5.8]):

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

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

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

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

$$\begin{aligned}
\frac{\binom{K}{l}}{\binom{L}{l}} &\leq \sqrt{\frac{8K(L-l)}{2\pi L(K-l)}} \exp\left(K\varphi(l/K) - L\varphi(l/L)\right) \\
&\leq \sqrt{\frac{8(1-l/L)}{2\pi(1-l/K)}} \exp\left(K\varphi(1-(3/2)\sqrt{p}) + c_1\mu p K \right. \\
&\qquad \left. - l\frac{1+2\mu p/\log(p)}{1-(3/2)\sqrt{p}}\varphi\left(\frac{1-(3/2)\sqrt{p}}{1+2\mu p/\log(p)}\right)\right) \\
&\leq \sqrt{\frac{8}{2\pi}} \exp\left(K\left(\varphi(1-(3/2)\sqrt{p}) + c_1\mu p \right.\right. \\
&\left.\left. - (1-(3/2)\sqrt{p} + \mu p/\log(p))\frac{1+2\mu p/\log(p)}{1-(3/2)\sqrt{p}}\varphi\left(\frac{1-(3/2)\sqrt{p}}{1+2\mu p/\log(p)}\right)\right)\right) \\
&= \sqrt{\frac{8}{2\pi}} \exp\left(K(F_\mu(p) + c_1\mu p)\right)
\end{aligned}$$

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

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

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

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

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

- We have

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

as $1 + 2\mu p/\log(p) \geq 1/2$. Also, $c \geq a \geq 1/8$ and $|\varphi'(x)| \leq -c_0 \log(1-x)$ for $x \in (1/8, 1)$. Therefore, $|\varphi(c) - \varphi(a)| \leq -c_0 \log(1-c)|c-a| \leq 4c_0 \log(1-c)\mu p/\log(p)$. 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, $|\varphi(c) - \varphi(a)| \leq c_4 \mu p$ 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 with respect to $C_3^{-1}/16$. □

C.2 Proof of Theorem 4.6

Upper bound on $t_{\lambda,d}(B)$ Let A, B be as in the statement of Theorem 4.6. A direct adaptation of [ALS13, Lemma 5] shows that for any $t \geq 0$, $h_d(B, t) \leq h_d(A, t+\gamma) + 2\gamma$. Therefore, according to Proposition 4.4, we have for $t^*(A) \leq t + \gamma < \tau(M)$,

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

Therefore, $h_d(t, B) < \lambda t$ 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 $\gamma \leq t^*(A) \leq \lambda^2 \tau(M)/24$ and if $t \in [t_0, t_1]$ with $t_0 = \frac{2t^*(A)}{\lambda}\left(1 + \frac{t^*(A)}{\tau(M)}\right) + \frac{6\gamma}{\lambda}$ and $t_1 = \frac{\tau(M)\lambda}{2}$. Therefore, $t_\lambda(B) \leq \frac{2t^*(A)}{\lambda}\left(1 + \frac{t^*(A)}{\tau(M)}\right) + \frac{6\gamma}{\lambda}$, as long as $t_{\max} < \tau(M)\lambda/2$ and $t_{\max} + \gamma < \tau(M)$.

Lower bound on $t_{\lambda,d}(A)$ in the noise-free case Assume that $\varepsilon(A) \leq \tau(M)/78$ so that Proposition 3.5 holds. Let $q \in M$ with $\varepsilon(A) = d(q, A)$. One has $q = \pi_M(x)$ for some $x \in \mathrm{Conv}_d(t^*(A); A)$, so that, by Proposition 3.5 and Lemma 3.3,

$$\begin{aligned}
d(x, A) \geq d(q, A) - \|x - q\| &\geq \frac{t^*(A)}{\left(1 + 6\frac{\varepsilon(A)}{\tau(M)}\right)} - \frac{t^*(A)^2}{\tau(M)} \\
&\geq t^*(A)\left(1 - 6\frac{\varepsilon(A)}{\tau(M)} - \frac{t^*(A)}{\tau(M)}\right) \\
&\geq t^*(A)\left(1 - 6\frac{2t^*(A)}{\tau(M)} - \frac{t^*(A)}{\tau(M)}\right) \geq t^*(A)\left(1 - 13\frac{t^*(A)}{\tau(M)}\right),
\end{aligned}$$

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

$$h_d(t^*(A), A) \geq t^*(A)\left(1 - 13\frac{t^*(A)}{\tau(M)}\right). \tag{C.16}$$

Therefore, if $\lambda \leq 1 - 13t^*(A)/\tau(M)$ and $t^*(A) < t_{\max}$, then $t_\lambda(A) \geq t^*(A)$.

Lower bound on $t_{\lambda,d}(A)$ in the tubular noise case [ALS13, Lemma 5] yields that for any $t \geq \gamma$,

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

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

$$h_d(B, t^*(A)+\gamma) \geq t^*(A)\left(1 - 13\frac{t^*(A)}{\tau(M)}\right) - 2\gamma. \tag{C.18}$$

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

$$13\frac{t^*(A)}{\tau(M)} \leq 1 - \lambda - (2+\lambda)\frac{\gamma}{t^*(A)}. \tag{C.19}$$

If $\gamma \leq (1-\lambda)\frac{t^*(A)}{6}$ and $13\frac{t^*(A)}{\tau(M)} \leq \frac{1-\lambda}{2}$, then (C.19) is satisfied, giving the desired lower bound on $t_\lambda(B)$ under those two conditions, should $t^*(A)+\gamma$ be smaller than $t_{\max}$.

C.3 Proof of Corollaries 4.7 and 4.9

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. As $M \subset \bigcup_{x\in A} \mathcal{B}_M(x, \varepsilon(A))$, 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. □

In the following proofs, we let $q = (\tau_{\min}, f_{\min}, +\infty, \eta) \in \mathcal{Q}^d$ and $P \in \mathcal{P}^d_{q,n}$. We write $\mathbb{X}_n = \{X_1, \dots, X_n\}$ a n-sample of law $\iota_\# P$ and let $\mathbb{Y}_n = \{Y_1, \dots, Y_n\}$ be the corresponding sample on $M := M(P)$. Also, $\gamma = \eta(\log n/n)^{2/d}$ and we fix $t_{\max} = 1/\log(n)$.

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

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

This relation holds 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 $\tau_{\min}$ and n is large enough. Note that the probability that $t^*(\mathbb{Y}_n)$ is smaller than $t_{\max}$ up to a constant is smaller than $C_{d,q}\exp(-nC'_{d,q}(\log n)^{-d})$ by

Proposition 3.6. Also, by Lemma C.6 and Proposition 3.5, Condition 3 is satisfied as long as n is large enough. Hence, if S denote the event that Conditions 1, 2 and 3 are satisfied for $B = \mathbb{X}_n$ and $A = \mathbb{Y}_n$, then, we have $\mathbb{P}(S) \geq 1 - C_{d,q} \exp(-C'_{d,q}(\log n)^{-d})$. Therefore, by using (C.21), the upper bound in Theorem 4.6, Proposition 3.5 and Lemma 3.6 (in that order), we obtain

$$\begin{aligned}
\mathbb{E}[d_H(\operatorname{Conv}_d(t_{\lambda,d}(\mathbb{X}_n);\mathbb{X}_n), M)] &\leq \mathbb{E}[d_H(\operatorname{Conv}_d(t_{\lambda,d}(\mathbb{X}_n);\mathbb{X}_n), M)\mathbf{1}\{S\}] + (\operatorname{diam}(M)+\gamma)\mathbb{P}(S^c) \\
&\leq \gamma + \frac{1}{\tau_{\min}}\mathbb{E}\left[\left(6\frac{t^*(\mathbb{Y}_n)}{\lambda} + \left(\frac{6}{\lambda}+1\right)\gamma\right)^2 \mathbf{1}\{S\}\right] + C_{d,q}\exp(-nC'_{d,q}(\log n)^{-d}) \\
&\leq \gamma + \frac{1}{\tau_{\min}}\frac{60}{\lambda^2}\mathbb{E}\left[\varepsilon(\mathbb{Y}_n)^2\right] + C_{d,q}\exp(-nC'_{d,q}(\log n)^{-d}) \\
&\leq \left(\frac{\log n}{n}\right)^{2/d}\left(\eta + \frac{1}{\tau_{\min}}\frac{121\pi^2}{\lambda^2(\omega_d f_{\min})^{2/d}}\right) \text{ if } n \text{ is large enough.}
\end{aligned}$$

As there exists a constant $C_{\kappa,\lambda}$ such that

$$\left(\eta + \frac{121\pi^2}{\lambda^2(\omega_d f_{\min})^{2/d}\tau_{\min}}\right) \leq C_{\kappa,\lambda}\left(\frac{\eta}{2} + \frac{C(1-\kappa)}{(\omega_d f_{\min})^{2/d}\tau_{\min}}\right)$$

for all $q \in \mathcal{Q}^d$, $1 \leq d \leq D$, the bound (4.8) follows from Theorem 2.4. □

Proof of Corollary 4.9. Recall that we assume that $\eta = 0$, so that $\mathbb{X}_n = \mathbb{Y}_n$. Theorem 3.2 in [BSW09] states that for $A \subset M$, if $t < \tau(M)/2$ and $t \geq 10\varepsilon(A)$, then

$$\angle(T_p(A,t), T_pM) \leq 6\frac{t}{\tau(M)}. \tag{C.22}$$

Assume that the conditions of Theorem 4.6 are satisfied for $\mathbb{X}_n$. Then, by Proposition 3.5 and Theorem 4.6,

$$\begin{aligned}
&11t_{\lambda,d}(\mathbb{X}_n) > 11t^*(\mathbb{X}_n) \geq 10\varepsilon(\mathbb{X}_n) \text{ and} \\
&11t_{\lambda,d}(\mathbb{X}_n) \leq \frac{2t^*(\mathbb{X}_n)}{\lambda}\left(1 + \frac{t^*(\mathbb{X}_n)}{\tau(M)}\right) \leq \frac{\lambda\tau(M)}{12}\left(1+\frac{1}{24}\right) < \frac{\tau(M)}{2}.
\end{aligned}$$

Hence, the upper bound in Theorem 4.6, Proposition 3.5 and Lemma 3.6 yield

$$\begin{aligned}
\mathbb{E}[\angle(T_pM, T_p(\mathbb{X}_n, 11t_\lambda(\mathbb{X}_n)))] &\leq \mathbb{E}[\angle(T_pM, T_p(\mathbb{X}_n, 11t_\lambda(\mathbb{X}_n)))\mathbf{1}\{S\}] + \mathbb{P}(S^c) \\
&\leq 66\frac{\mathbb{E}t_{\lambda,d}(\mathbb{X}_n)}{\tau_{\min}} + C_{q,d}\exp(-C'_{q,d}(\log n)^{-d}) \leq \frac{136}{\lambda}\frac{\mathbb{E}\varepsilon(\mathbb{X}_n)}{\tau_{\min}} + C_{q,d}\exp(-nC'_{d,q}(\log n)^{-d}) \\
&\leq \frac{136}{\lambda}\frac{\sqrt{\mathbb{E}\varepsilon(\mathbb{X}_n)^2}}{\tau_{\min}} + C_{q,d}\exp(-C'_{q,d}(\log n)^{-d}) \leq \frac{137\sqrt{2}\pi}{\lambda\tau_{\min}(\omega_d f_{\min})^{1/d}}\left(\frac{\log n}{n}\right)^{1/d}
\end{aligned}$$

if n is large enough. □

D Precise lower bound on the minimax risk

The goal of this section is to show the lower bound in Theorem 2.4. To do so, we adapt the construction made in [KZ15] so that the lower bound holds with an explicit constant. Let $0 < d < D$ and $q = (\tau_{\min}, f_{\min}, f_{\max}, \eta) \in \mathcal{Q}^d(\kappa)$. We denote by $M(P)$ the underlying manifold of $P \in \mathcal{P}(\kappa)$. The lowerbound is based on Le Cam's lemma:

Lemma D.1. *Let $\mathcal{P}^{(1)}$, $\mathcal{P}^{(2)}$ be two subfamilies of $\mathcal{P}^d_{q,n}$ which are ε-separated, in the sense that $d_H(M(P^{(1)}), M(P^{(2)})) \geq 2\varepsilon$ for all $P^{(1)} \in \mathcal{P}^{(1)}$, $P^{(2)} \in \mathcal{P}^{(2)}$. Then,*

$$m_n(M, \mathcal{P}^d_{q,n}) \geq \varepsilon \left| \left(\frac{1}{\#\mathcal{P}^{(1)}} \sum_{P^{(1)} \in \mathcal{P}^{(1)}} \iota_\# P^{(1)} \right) \wedge \left(\frac{1}{\#\mathcal{P}^{(2)}} \sum_{P^{(2)} \in \mathcal{P}^{(2)}} \iota_\# P^{(2)} \right) \right|, \tag{D.1}$$

where $|P \wedge Q|$ is the testing affinity between two distributions P and Q and $\iota : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^D$ is the addition.

To obtain a lowerbound on the minimax risk, authors in [KZ15] 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 $\mathcal{P}^d_{q,n}$ 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 $f_{\min}\omega_d\tau^d_{\min} \leq \kappa < 1$). 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 $M_0 \subset \mathbb{R}^{d+1} \subset \mathbb{R}^D$ 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^ε_s by letting for $x \in \mathbb{R}^D$

$$\Phi^\varepsilon_s(x) = x \left(1 + \frac{\varepsilon}{R} \sum_{y \in A} s(y) \phi\left(\frac{\|x - y\|}{\delta} \right) \right). \tag{D.2}$$

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

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

diffeomorphism on its image, with

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

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

$$d_x\Phi_s^\varepsilon(h) = h + h\frac{\varepsilon}{R}\sum_{y\in A} s(y)\phi\left(\frac{|x-y|}{\delta}\right) + x\frac{\varepsilon}{R}\sum_{y\in A}\frac{1}{\delta}s(y)\phi'\left(\frac{|x-y|}{\delta}\right)\frac{\langle x-y, h\rangle}{|x-y|}. \tag{D.4}$$

Hence,

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

where $c_1 = \|\phi\|_\infty + 3\|\phi'\|_\infty$. A similar computation gives that $\|d_x^2\Phi_s^\varepsilon\|_{\mathrm{op}} \le c_2\varepsilon/\delta^2$ for $c_2 = 2\|\phi'\|_\infty + 3/c_1(\|\phi'\|_\infty + \|\phi''\|_\infty)$. We eventually show the injectivity: if $\Phi_s^\varepsilon(x) = \Phi_s^\varepsilon(x')$, then x and x' are colinear. Also, if $c_1\varepsilon/\delta < 1$, one can check using (D.4) that the derivative of the function $r \mapsto \langle\Phi_s^\varepsilon(ru), u\rangle$ for u an unit vector is increasing, proving the injectivity. □

Therefore, from [Fed59, Theorem 14.19], we infer that $M_s^\varepsilon := \Phi_s^\varepsilon(M)$ is a manifold with reach larger than

$$\tau(M_s^\varepsilon) \ge R\min\left(1 - c_1\varepsilon/\delta, \frac{(1-c_1\varepsilon/\delta)^2}{1 + c_1\varepsilon/\delta + Rc_2\varepsilon/\delta^2}\right). \tag{D.5}$$

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

$$\begin{aligned}\mathrm{Vol}(M_s^\varepsilon) &= \int_{M_0} J\Phi_s^\varepsilon(x)\mathrm{d}x = \omega_d R^d + \sum_{y\in A}\int_{\mathcal{B}_{M_0}(y,2\delta)}(J\Phi_s^\varepsilon(x) - 1)\mathrm{d}x \\ &\le \omega_d R^d + 2mC_dc_1\frac{\varepsilon}{\delta}\mathrm{Vol}(\mathcal{B}_{M_0}(y, 2\delta)) \le \omega_d R^d\left(1 + C_dc_1\frac{\varepsilon}{\delta}\right),\end{aligned} \tag{D.6}$$

where we used that $\det(N) - 1 \le C_d\|N - \mathrm{Id}\|_{\mathrm{op}}$ for some constant C_d if N is a matrix of size d with operator norm smaller than 1, the fact that $2m\,\mathrm{Vol}(\mathcal{B}_{M_0}(y, 2\delta)) \le \mathrm{Vol}(M_0)$, and Lemma D.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 ν is chosen so that $R > \tau_{\min}(1 + c_2/\nu^2)$, say $\nu^2 = \frac{2c_2\tau_{\min}}{R-\tau_{\min}}$. Then, if ε/δ is small enough, we have $\mathrm{Vol}(M_s^\varepsilon) \le 1/f_{\min}$ and $\tau(M_s^\varepsilon) \ge \tau_{\min}$ by (D.6) and (D.5). 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

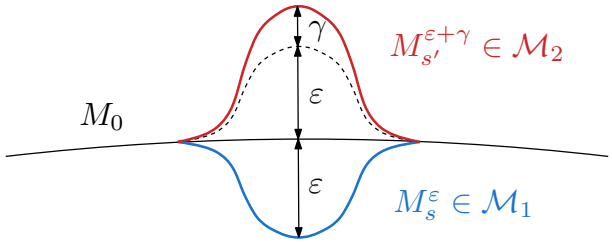


Figure 5 – An element $P^{(1)} \in \mathcal{P}^{(1)}$ has its first marginal supported on the blue manifold M_s^ε (lower bump), whereas an element $P^{(2)} \in \mathcal{P}^{(2)}$ is such that $P_1^{(2)}$ is supported on the red manifold $M_{s'}^{\varepsilon+\gamma}$ (upper bump) and $\iota_\# P^{(2)}$ is the uniform distribution on the dotted manifold.

$(Q_s^\varepsilon, \delta_0)$ 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 $\mathcal{P}_{q,n}^d$. We then define $\mathcal{P}^{(2)}$ as follows: let $X \sim Q_s^\varepsilon$ where Q_s^ε is the uniform distribution on a manifold of $M_s^\varepsilon \in \mathcal{M}^{(2)}$. Then, we have $X = \Phi_s^\varepsilon(V)$ for some $V \in M_0$, and we let

$$Y = \Phi_s^{\varepsilon+\gamma}(V), \qquad Z = X - Y.$$

An element of $\mathcal{P}^{(2)}$ is then given by the law of the couple (Y, Z). Note that for $P^{(2)} \in \mathcal{P}^{(2)}$, $\iota_\# P^{(2)}$ is the uniform distribution on a manifold of $\mathcal{M}^{(2)}$. Also, $M(P^{(2)})$ is equal to $M_s^{\varepsilon+\gamma} = \Phi_s^{\varepsilon+\gamma} \circ (\Phi_s^\varepsilon)^{-1}(M_s^\varepsilon)$ for some $M_s^\varepsilon \in \mathcal{M}^{(2)}$. By (D.6) and (D.5), its reach is also larger than $\tau_{\min}$, and its volume is smaller than $1/f_{\min}$ if $(\varepsilon+\gamma)/\delta$ is small enough. Note also that $|Z| = |\Phi_s^\varepsilon(V) - \Phi_s^{\varepsilon+\gamma}(V)| \le |V|\gamma/R \le \gamma$. Hence, $\mathcal{P}^{(2)}$ is indeed a subset of $\mathcal{P}_{q,n}^d$.

By construction, the two families $\mathcal{P}^{(1)}$, $\mathcal{P}^{(2)}$ are $(2\varepsilon+\gamma)$-separated (see Figure 5). Hence, we can apply Le Cam's lemma. The exact same computations than in [KZ15, Section 3] show that the testing affinity between $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ converge to 1 as long as $4m = n/\log n$. Thus, Le Cam's Lemma (D.1) yields

$$\liminf_n \frac{m_n(M, \mathcal{P}_{q,n}^d)}{\left(\frac{\log n}{n}\right)^{2/d}} \ge \liminf_n \left((m/4)^{2/d}\varepsilon + \frac{\eta}{2} \right). \tag{D.7}$$

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

$$\begin{aligned}\liminf_n \frac{m_n(M, \mathcal{P}_{q,n}^d)}{\left(\frac{\log n}{n}\right)^{2/d}} &\ge \frac{c_0^2}{8^{2/d}} \frac{R^2}{\delta^2}\varepsilon + \frac{\eta}{2} = \frac{c_0^2}{8^{2/d}} \frac{R}{\nu^2} + \frac{\eta}{2} \\ &= \frac{c_0^2}{8^{2/d}} \frac{R(R-\tau_{\min})}{2c_2\tau_{\min}} + \frac{\eta}{2} \ge \frac{c_3}{(\omega_d f_{\min})^{1/d}\tau_{\min}} \left(\frac{1}{(\omega_d f_{\min})^{1/d}} - \tau_{\min} \right) + \frac{\eta}{2},\end{aligned}$$

for some absolute constant c_3, where we used that $R - \tau_{\min} = \frac{1}{2}\left(\frac{1}{(\omega_d f_{\min})^{1/d}} - \tau_{\min} \right)$ by definition and that $R \ge \frac{1}{2}(\omega_d f_{\min})^{-1/d}$. As $\tau_{\min} \le \kappa/(\omega_d f_{\min})^{1/d}$, we obtain the conclusion.