NON-ASYMPTOTIC RATES FOR MANIFOLD, TANGENT SPACE AND CURVATURE ESTIMATION

By Eddie Aamari[§,*,†,‡] and Clément Levrard[¶,*,†]

U.C. San Diego[§] , *Université Paris-Diderot*[¶]

Abstract: Given a noisy sample from a submanifold $M \subset \mathbb{R}^D$, we derive optimal rates for the estimation of tangent spaces $T_X M$, the second fundamental form II_X^M, and the submanifold M. After motivating their study, we introduce a quantitative class of $\mathcal{C}^k$-submanifolds in analogy with Hölder classes. The proposed estimators are based on local polynomials and allow to deal simultaneously with the three problems at stake. Minimax lower bounds are derived using a conditional version of Assouad's lemma when the base point X is random.

1. Introduction. A wide variety of data can be thought of as being generated on a shape of low dimensionality compared to possibly high ambient dimension. This point of view led to the development of the so-called topological data analysis, which proved fruitful for instance when dealing with physical parameters subject to constraints, biomolecule conformations, or natural images [29]. This field intends to associate geometric quantities to data without regard of any specific coordinate system or parametrization. If the underlying structure is sufficiently smooth, one can model a point cloud $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ as being sampled on a d-dimensional submanifold $M \subset \mathbb{R}^D$. In such a case, geometric and topological intrinsic quantities include (but are not limited to) homology groups [22], persistent homology [10], volume [4], differential quantities [6] or the submanifold itself [14, 1, 20].

The present paper focuses on optimal rates for estimation of quantities up to order two: (0) the submanifold itself, (1) tangent spaces, and (2) second fundamental forms.

Among these three questions, a special attention has been paid to the estimation of the submanifold. In particular, it is a central problem in manifold learning. Indeed, there exists a wide bunch of algorithms intended to reconstruct submanifolds from point clouds (Isomap [26], LLE [23], and restricted Delaunay Complexes [5, 8] for instance), but few come with theoretical guarantees [14, 1, 20]. Up to our knowledge, minimax lower bounds

[*] Research supported by ANR project TopData ANR-13-BS01-0008

[†] Research supported by Advanced Grant of the European Research Council GUDHI

[‡] Supported by the Conseil régional d'Île-de-France program RDM-IdF

were used to prove optimality in only one case [14]. Some of these reconstruction procedures are based on tangent space estimation [5, 1, 8]. Tangent space estimation itself also yields interesting applications in manifold clustering [13, 3]. Estimation of curvature-related quantities naturally arises in shape reconstruction, since curvature can drive the size of a meshing. As a consequence, most of the associated results deal with the case $d = 2$ and $D = 3$, though some of them may be extended to higher dimensions [21, 17]. Several algorithms have been proposed in that case [24, 6, 21, 17], but with no analysis of their performances from a statistical point of view.

To assess the quality of such a geometric estimator, the class of submanifolds over which the procedure is evaluated has to be specified. Up to now, the most commonly used model for submanifolds relied on the reach τ_M, a generalized convexity parameter. Assuming $\tau_M \geq \tau_{min} > 0$ involves both local regularity — a bound on curvature — and global regularity — no arbitrarily pinched area —. This $\mathcal{C}^2$-like assumption has been extensively used in the computational geometry and geometric inference fields [1, 22, 10, 4, 14]. One attempt of a specific investigation for higher orders of regularity $k \geq 3$ has been proposed in [6].

Many works suggest that the regularity of the submanifold has an important impact on convergence rates. This is pretty clear for tangent space estimation, where convergence rates of PCA-based estimators range from $(1/n)^{1/d}$ in the $\mathcal{C}^2$ case [1] to $(1/n)^{\alpha}$ with $1/d < \alpha < 2/d$ in more regular settings [25, 27]. In addition, it seems that PCA-based estimators are outperformed by estimators taking into account higher orders of smoothness [7, 6], for regularities at least $\mathcal{C}^3$. For instance fitting quadratic terms leads to a convergence rate of order $(1/n)^{2/d}$ in [7]. These remarks naturally led us to investigate the properties of local polynomial approximation for regular submanifolds, where "regular" has to be properly defined. Local polynomial fitting for geometric inference was studied in several frameworks such as [6]. In some sense, a part of our work extends these results, by investigating the dependency of convergence rates on the sample size n, but also on the order of regularity k and the ambient and intrinsic dimensions d and D.

1.1. *Overview of the Main Results.* In this paper, we build a collection of models for $\mathcal{C}^k$-submanifolds ($k \geq 3$) that naturally generalize the commonly used one for $k = 2$ (Section 2). Roughly speaking, these models are defined by their local differential regularity k in the usual sense, and by their minimum reach $\tau_{min} > 0$ that may be thought of as a global regularity parameter (see Section 2.2). On these models, we study the non-asymptotic rates of estimation for tangent space, curvature, and manifold estimation (Section 3).

Roughly speaking, if M is a $\mathcal{C}^k_{\tau_{min}}$ submanifold and if $Y_1, \ldots, Y_n$ is an n-sample drawn on M uniformly enough, then we can derive the following minimax bounds:

$$\textit{(Theorems 2 and 3)} \qquad \inf_{\hat{T}} \sup_{\substack{M \in \mathcal{C}^k \\ \tau_M \geq \tau_{min}}} \mathbb{E} \max_{1 \leq j \leq n} \angle\big(T_{Y_j} M, \hat{T}_j\big) \asymp \left(\frac{1}{n}\right)^{\frac{k-1}{d}},$$

where $T_y M$ denotes the tangent space of M at y;

$$\textit{(Theorems 4 and 5)} \qquad \inf_{\widehat{II}} \sup_{\substack{M \in \mathcal{C}^k \\ \tau_M \geq \tau_{min}}} \mathbb{E} \max_{1 \leq j \leq n} \big\| II^M_{Y_j} - \widehat{II}_j \big\| \asymp \left(\frac{1}{n}\right)^{\frac{k-2}{d}},$$

where II^M_y denotes the second fundamental form of M at y;

$$\textit{(Theorems 6 and 7)} \qquad \inf_{\hat{M}} \sup_{\substack{M \in \mathcal{C}^k \\ \tau_M \geq \tau_{min}}} \mathbb{E}\ d_H\big(M, \hat{M}\big) \asymp \left(\frac{1}{n}\right)^{\frac{k}{d}},$$

where d_H denotes the Hausdorff distance.

These results shed light on the influence of k, d, and n on these estimation problems, showing for instance that the ambient dimension D plays no role. The estimators proposed for the upper bounds all rely on the analysis of local polynomials, and allow to deal with the three estimation problems in a unified way (Section 5.1). Some of the lower bounds are derived using a new version of Assouad's Lemma (Section 5.2.2).

We also emphasize the influence of the reach τ_M of the manifold M in Theorem 1. Indeed, we show that whatever the local regularity k of M, if we only require $\tau_M \geq 0$, then for any fixed point $y \in M$,

$$\inf_{\hat{T}} \sup_{\substack{M \in \mathcal{C}^k \\ \tau_M \geq 0}} \mathbb{E} \angle\big(T_y M, \hat{T}\big) \geq 1/2, \qquad \inf_{\widehat{II}} \sup_{\substack{M \in \mathcal{C}^k \\ \tau_M \geq 0}} \mathbb{E} \big\| II^M_y - \widehat{II} \big\| \geq c > 0,$$

assessing that the global regularity parameter $\tau_{min} > 0$ is crucial for estimation purpose.

It is worth mentioning that our bounds also allow for perpendicular noise of amplitude $\sigma > 0$. When $\sigma \lesssim (1/n)^{\alpha/d}$ for $1 \leq \alpha$, then our estimators behave as if the corrupted sample $X_1, \ldots, X_n$ were exactly drawn on a manifold with regularity α. Hence our estimators turn out to be optimal whenever $\alpha \geq k$. If $\alpha < k$, the lower bounds suggest that better rates could be obtained with different estimators, by pre-processing data as in [15] for instance.

For the sake of completeness, geometric background and proofs of technical lemmas are given in the Appendix.

2. $\mathcal{C}^k$ Models for Submanifolds.

2.1. *Notation.* Throughout the paper, we consider d-dimensional compact submanifolds $M \subset \mathbb{R}^D$ without boundary. The submanifolds will always be assumed to be at least $\mathcal{C}^2$. For all $p \in M$, T_pM stands for the tangent space of M at p [9, Chapter 0]. We let $II_p^M : T_pM \times T_pM \to T_pM^{\perp}$ denote the second fundamental form of M at p [9, p. 125]. II_p^M characterizes the curvature of M at p. The standard inner product in $\mathbb{R}^D$ is denoted by $\langle \cdot, \cdot \rangle$ and the Euclidean distance by $\|\cdot\|$. Given a linear subspace $T \subset \mathbb{R}^D$, write $T^{\perp}$ for its orthogonal space. We write $\mathcal{B}(p,r)$ for the closed Euclidean ball of radius $r > 0$ centered at $p \in \mathbb{R}^D$, and for short $\mathcal{B}_T(p,r) = \mathcal{B}(p,r) \cap T$. For a smooth function $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ and $i \geq 1$, we let $d_x^i \Phi$ denote the ith order differential of Φ at $x \in \mathbb{R}^D$. For a linear map A defined on $T \subset \mathbb{R}^D$, $\|A\|_{\mathrm{op}} = \sup_{v \in T} \frac{\|Av\|}{\|v\|}$ stands for the operator norm. We adopt the same notation $\|\cdot\|_{op}$ for tensors, i.e. multilinear maps. Similarly, if $\{A_x\}_{x \in T'}$ is a family of linear maps, its L^{∞} operator norm is denoted by $\|A\|_{op} = \sup_{x \in T'} \|A_x\|_{op}$. When it is well defined, we will write $\pi_B(z)$ for the projection of $z \in \mathbb{R}^D$ onto the closed subset $B \subset \mathbb{R}^D$, that is the nearest neighbor of z in B. The distance between two linear subspaces $U, V \subset \mathbb{R}^D$ of the same dimension is measured by the principal angle $\angle(U,V) = \|\pi_U - \pi_V\|_{\mathrm{op}}$. The Hausdorff distance [14] in $\mathbb{R}^D$ is denoted by d_H. For a probability distribution P, $\mathbb{E}_P$ stands for the expectation with respect to P. We write $P^{\otimes n}$ for the n-times tensor product of P.

Throughout this paper, C_α will denote a generic constant depending on the parameter α. For clarity's sake, C'_α, c_α, or c'_α may also be used when several constants are involved.

2.2. *Reach and Regularity of Submanifolds.* As introduced in [11], the reach τ_M of a subset $M \subset \mathbb{R}^D$ is the maximal neighborhood radius for which the projection π_M onto M is well defined. More precisely, denoting by $d(\cdot, M)$ the distance to M, the medial axis of M is defined to be the set of points which have at least two nearest neighbors on M, that is

$$Med(M) = \left\{ z \in \mathbb{R}^D | \exists p \neq q \in M, \|z-p\| = \|z-q\| = d(z,M) \right\}.$$

The reach is then defined by

$$\tau_M = \inf_{p \in M} d\left(p, Med(M)\right) = \inf_{z \in Med(M)} d\left(z, M\right).$$

It gives a minimal scale of geometric and topological features of M. As a generalized convexity parameter, τ_M is a key parameter in reconstruction

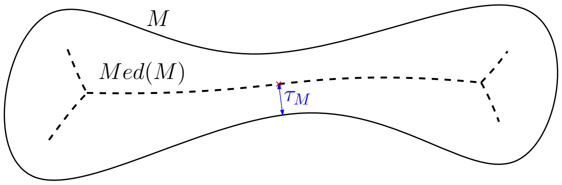


Figure 1: Medial axis and reach of a closed curve in the plane.

[1, 14] and in topological inference [22]. Having $\tau_M \geq \tau_{min} > 0$ prevents M from almost auto-intersecting, and bounds its curvature in the sense that $\left\| II_p^M \right\|_{op} \leq \tau_M^{-1} \leq \tau_{min}^{-1}$ for all $p \in M$ [22, Proposition 6.1].

For $\tau_{min} > 0$, we let $\mathcal{C}^2_{\tau_{min}}$ denote the set of d-dimensional compact connected submanifolds M of $\mathbb{R}^D$ such that $\tau_M \geq \tau_{min} > 0$. A key property of submanifolds $M \in \mathcal{C}^2_{\tau_{min}}$ is the existence of a parametrization closely related to the projection onto tangent spaces. We let $\exp_p : T_pM \to M$ denote the geodesic map of M [9, Chapter 3], that is defined by $\exp_p(v) = \gamma_{p,v}(1)$, where $\gamma_{p,v}$ is the unique constant speed geodesic path of M with initial value p and velocity v.

Lemma 1. *If $M \in \mathcal{C}^2_{\tau_{min}}$, $\exp_p : \mathcal{B}_{T_pM}(0, \tau_{min}/4) \to M$ is one-to-one. Moreover, it can be written as*

$$\begin{aligned} \exp_p \colon \mathcal{B}_{T_pM}(0, \tau_{min}/4) &\longrightarrow M \\ v &\longmapsto p + v + \mathbf{N}_p(v) \end{aligned}$$

with $\mathbf{N}_p$ such that for all $v \in \mathcal{B}_{T_pM}(0, \tau_{min}/4)$,

$$\mathbf{N}_p(0) = 0, \quad d_0\mathbf{N}_p = 0, \quad \left\| d_v\mathbf{N}_p \right\|_{op} \leq L_\perp \left\| v \right\|,$$

where $L_\perp = 5/(4\tau_{min})$. Furthermore, for all $p, y \in M$,

$$y - p = \pi_{T_pM}(y - p) + R_2(y - p),$$

where $\left\| R_2(y-p) \right\| \leq \frac{\|y-p\|^2}{2\tau_{min}}$.

A proof of Lemma 1 is given in Section A.1 of the Appendix. In other words, elements of $\mathcal{C}^2_{\tau_{min}}$ have local parametrizations on top of their tangent spaces that are defined on neighborhoods with a minimal radius, and these parametrizations differ from the identity map by at most a quadratic term. The existence of such local parametrizations leads to the following convergence result: if data $Y_1, \ldots, Y_n$ are drawn uniformly enough on $M \in \mathcal{C}^2_{\tau_{min}}$, then it is shown in [1, Proposition 14] that a tangent space estimator $\hat{T}$ based on local PCA achieves

$$\mathbb{E} \max_{1 \leq j \leq n} \angle\big(T_{Y_j} M, \hat{T}_j\big) \leq C \left(\frac{1}{n}\right)^{\frac{1}{d}}.$$

When M is smoother, it has been proved in [7] that a convergence rate in $n^{-2/d}$ might be achieved, based on the existence of a local order 3 Taylor expansion of the submanifold on top of its tangent spaces. Thus, a natural extension of the $\mathcal{C}^2_{\tau_{min}}$ model to $\mathcal{C}^k$-submanifolds should ensure that such an expansion exists at order k and satisfies some regularity constraints. To this aim, we introduce the following class of regularity $\mathcal{C}^k_{\tau_{min},\mathbf{L}}$.

Definition 1. *For $k \geq 3$, $\tau_{min} > 0$, and $\mathbf{L} = (L_\perp, L_3, \ldots, L_k)$, we let $\mathcal{C}^k_{\tau_{min},\mathbf{L}}$ denote the set of d-dimensional compact connected submanifolds M of $\mathbb{R}^D$ with $\tau_M \geq \tau_{min}$ and such that, for all $p \in M$, there exists a local one-to-one parametrization Ψ_p of the form:*

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

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

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

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

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

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

The Ψ_p's can be seen as unit parametrizations of M. The conditions on $\mathbf{N}_p(0)$, $d_0\mathbf{N}_p$, and $d_v^2\mathbf{N}_p$ ensure that Ψ_p^{-1} is close to the projection π_{T_pM}. The bounds on $d_v^i\mathbf{N}_p$ $(3 \leq i \leq k)$ allow to control the coefficients of the polynomial expansion we seek. Indeed, whenever $M \in \mathcal{C}^k_{\tau_{min},\mathbf{L}}$, Lemma 2 shows that for every p in M, and y in $\mathcal{B}\big(p, \frac{\tau_{min} \wedge L_\perp^{-1}}{4}\big) \cap M$,

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

where π^* denotes the orthogonal projection onto T_pM, the T_i^* are i-linear maps from T_pM to $\mathbb{R}^D$ with $\|T_i^*\|_{op} \leq L_i'$ and R_k satisfies $\|R_k(y-p)\| \leq C\|y-p\|^k$, where the constants C and the L_i''s depend on the parameters τ_{min}, d, k, $L_\perp, \ldots, L_k$.

Note that for $k \geq 3$ the exponential map can happen to be only $\mathcal{C}^{k-2}$ for a $\mathcal{C}^k$-submanifold [18]. Hence, it may not be a good choice of Ψ_p. However, for $k=2$, taking $\Psi_p = \exp_p$ is sufficient for our purpose. For ease of notation, we may write $\mathcal{C}^2_{\tau_{min},\mathbf{L}}$ although the specification of $\mathbf{L}$ is useless. In this case, we implicitly set by default $\Psi_p = \exp_p$ and $L_\perp = 5/(4\tau_{min})$. As will be shown in Theorem 1, the global assumption $\tau_M \geq \tau_{min} > 0$ cannot be dropped, even when higher order regularity bounds L_i's are fixed.

Let us now describe the statistical model. Every d-dimensional submanifold $M \subset \mathbb{R}^D$ inherits a natural uniform volume measure by restriction of the ambient d-dimensional Hausdorff measure $\mathcal{H}^d$. In what follows, we will consider probability distributions that are almost uniform on some M in $\mathcal{C}^k_{\tau_{min},\mathbf{L}}$, with some bounded noise, as stated below.

Definition 2 (Noise-Free and Tubular Noise Models).
- *(Noise-Free Model) For $k \geq 2$, $\tau_{min} > 0$, $\mathbf{L} = (L_\perp, L_3, \ldots, L_k)$ and $f_{min} \leq f_{max}$, we let $\mathcal{P}^k_{\tau_{min},\mathbf{L},f_{min},f_{max}}$ denote the set of distributions P_0 with support $M \in \mathcal{C}^k_{\tau_{min},\mathbf{L}}$ that have a density f with respect to the volume measure on M, and such that for all $y \in M$,*

$$0 < f_{min} \leq f(y) \leq f_{max} < \infty.$$

- *(Tubular Noise Model) For $0 \leq \sigma < \tau_{min}$, we denote by $\mathcal{P}^k_{\tau_{min},\mathbf{L},f_{min},f_{max}}(\sigma)$ the set of distributions of random variables $X = Y + Z$, where Y has distribution $P_0 \in \mathcal{P}^k_{\tau_{min},\mathbf{L},f_{min},f_{max}}$, and $Z \in T_YM^\perp$ with $\|Z\| \leq \sigma$ and $\mathbb{E}(Z|Y) = 0$.*

For short, we write $\mathcal{P}^k$ and $\mathcal{P}^k(\sigma)$ when there is no ambiguity. We denote by $\mathbb{X}_n$ an i.i.d. n-sample $\{X_1, \ldots, X_n\}$, that is, a sample with distribution $P^{\otimes n}$ for some $P \in \mathcal{P}^k(\sigma)$, so that $X_i = Y_i + Z_i$, where Y has distribution $P_0 \in \mathcal{P}^k$, $Z \in \mathcal{B}_{T_YM^\perp}(0,\sigma)$ with $\mathbb{E}(Z|Y) = 0$. It is immediate that for $\sigma < \tau_{min}$, we have $Y = \pi_M(X)$. Note that the tubular noise model $\mathcal{P}^k(\sigma)$ is a slight generalization of that in [15].

In what follows, though M is unknown, all the parameters of the model will be assumed to be known, including the intrinsic dimension d and the order of regularity k. We will also denote by $\mathcal{P}^k_{(x)}$ the subset of elements in $\mathcal{P}^k$ whose support contains a prescribed $x \in \mathbb{R}^D$.

In view of our minimax study on $\mathcal{P}^k$, it is important to ensure by now that $\mathcal{P}^k$ is stable with respect to deformations and dilations.

Proposition 1. *Let $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ be a global $\mathcal{C}^k$-diffeomorphism. If $\left\| d\Phi - I_D \right\|_{op}$, $\left\| d^2\Phi \right\|_{op}$, ..., $\left\| d^k\Phi \right\|_{op}$ are small enough, then for all P in $\mathcal{P}^k_{\tau_{min},\mathbf{L},f_{min},f_{max}}$, the pushforward distribution $P' = \Phi_* P$ belongs to $\mathcal{P}^k_{\tau_{min}/2,2\mathbf{L},f_{min}/2,2f_{max}}$.*

Moreover, if $\Phi = \lambda I_D$ $(\lambda > 0)$ is an homogeneous dilation, then $P' \in \mathcal{P}^k_{\lambda\tau_{min},\mathbf{L}_{(\lambda)},f_{min}/\lambda^d,f_{max}/\lambda^d}$, where $\mathbf{L}_{(\lambda)} = (L_\perp/\lambda, L_3/\lambda^2, \ldots, L_k/\lambda^{k-1})$.

Proposition 1 follows from a geometric reparametrization argument (Proposition A.5 in the Appendix) and a change of variable result for the Hausdorff measure (Lemma A.6 in the Appendix).

2.3. *Necessity of a Global Assumption.* In the previous Section 2.2, we generalized $\mathcal{C}^2$-like models — stated in terms of reach — to $\mathcal{C}^k$, for $k \geq 3$, by imposing higher order differentiability bounds on parametrizations Ψ_p's. The following Theorem 1 shows that the global assumption $\tau_M \geq \tau_{min} > 0$ is necessary for estimation purpose.

Theorem 1. *Assume that $\tau_{min} = 0$. If $D \geq d+3$, then for all $k \geq 3$ and $L_\perp > 0$, provided that $L_3/L_\perp^2, \ldots, L_k/L_\perp^{k-1}, L_\perp^d/f_{min}$ and $f_{max}/L_\perp^d$ are large enough (depending only on d and k), for all $n \geq 1$,*

$$\inf_{\hat{T}} \sup_{P \in \mathcal{P}^k_{(x)}} \mathbb{E}_{P^{\otimes n}} \angle\big(T_x M, \hat{T}\big) \geq \frac{1}{2} > 0,$$

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

Moreover, for any $D \geq d+1$, provided that $L_3/L_\perp^2, \ldots, L_k/L_\perp^{k-1}, L_\perp^d/f_{min}$ and $f_{max}/L_\perp^d$ are large enough (depending only on d and k), for all $n \geq 1$,

$$\inf_{\widehat{II}} \sup_{P \in \mathcal{P}^k_{(x)}} \mathbb{E}_{P^{\otimes n}} \left\| II_x^M \circ \pi_{T_x M} - \widehat{II} \right\|_{op} \geq \frac{L_\perp}{4} > 0,$$

where the infimum is taken over all the estimators $\widehat{II} = \widehat{II}\big(X_1, \ldots, X_n\big)$.

The proof of Theorem 1 can be found in Section C.5 of the Appendix. In other words, if the class of submanifolds is allowed to have arbitrarily small reach, no estimator can perform uniformly well to estimate neither $T_x M$ nor II_x^M. And this, even though each of the underlying submanifolds have arbitrarily smooth parametrizations. Indeed, if two parts of M can nearly intersect around x at an arbitrarily small scale $\Lambda \to 0$, no estimator can decide whether the direction (resp. curvature) of M at x is that of the first part or the second part (see Figures 2 and 3).

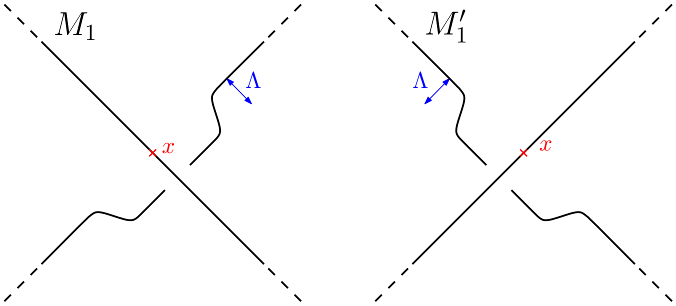


Figure 2: Inconsistency of tangent space estimation for $\tau_{min} = 0$.

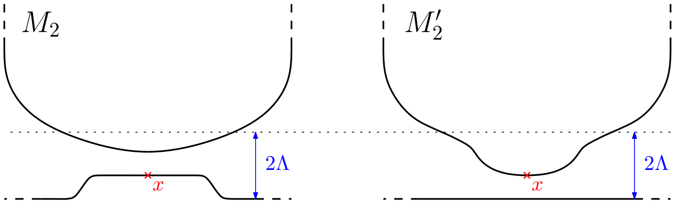


Figure 3: Inconsistency of curvature estimation for $\tau_{min} = 0$.

3. Main Results.

Let us now move to the statement of the main results. Given an i.i.d. n-sample $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ with unknown common distribution $P \in \mathcal{P}^k(\sigma)$, we detail non-asymptotic rates for the estimation of tangent spaces $T_{Y_j}M$, second fundamental forms $II^M_{Y_j}$, and M itself.

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

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

where π ranges among all the orthogonal projectors on d-dimensional subspaces, and $T_i : \left(\mathbb{R}^D\right)^i \to \mathbb{R}^D$ among the symmetric tensors of order i such that $\|T_i\|_{op} \leq t$. For $k = 2$, the sum over the tensors T_i is empty, and the integrated term reduces to $\|x - \pi(x)\|^2 \mathbb{1}_{\mathcal{B}(0,h)}(x)$. By compactness of the domain of minimization, such a minimizer exists almost surely. In what follows, we will work with a maximum scale $h \leq h_0$, with

$$h_0 = \frac{\tau_{min} \wedge L_\perp^{-1}}{8}.$$

The set of d-dimensional orthogonal projectors is not convex, which leads to a more involved optimization problem than usual least squares. In practice, this problem may be solved using tools from optimization on Grassman manifolds [28], or adopting a two-stage procedure such as in [6]: from local PCA, a first d-dimensional space is estimated at each sample point, along with an orthonormal basis of it. Then, the optimization problem (2) is expressed as a minimization problem in terms of the coefficients of $(\pi_j, T_{2,j}, \ldots, T_{k,j})$ in this basis under orthogonality constraints. It is worth mentioning that a similar problem is explicitly solved in [7], leading to an optimal tangent space estimation procedure in the case $k = 3$.

The constraint $\|T_i\|_{op} \leq t$ involves a parameter t to be calibrated. As will be shown in the following section, it is enough to choose t roughly smaller than $1/h$, but still larger than the unknown norm of the optimal tensors $\|T_i^*\|_{op}$. Hence, for $h \to 0$, the choice $t = h^{-1}$ works to guarantee optimal convergence rates. Such a constraint on the higher order tensors might have been stated under the form of a $\|.\|_{op}$-penalized least squares minimization — as in ridge regression — leading to the same results.

3.1. *Tangent Spaces.* By definition, the tangent space $T_{Y_j}M$ is the best linear approximation of M nearby Y_j. Thus, it is natural to take the range of the first order term minimizing (2) and write $\hat{T}_j = \operatorname{im} \hat{\pi}_j$. The $\hat{T}_j$'s approximate simultaneously the $T_{Y_j}M$'s with high probability, as stated below.

THEOREM 2. *Assume that $t \geq C_{k,d,\tau_{min},\mathbf{L}} \geq \sup_{2 \leq i \leq k} \|T_i^*\|_{op}$. Set $h = \left(C_{d,k} \frac{f_{max}^2 \log n}{f_{min}^3 (n-1)}\right)^{1/d}$, for $C_{d,k}$ large enough, and assume that $\sigma \leq h/4$. If n is large enough so that $h \leq h_0$, then with probability at least $1 - \left(\frac{1}{n}\right)^{k/d}$,*

$$\max_{1 \leq j \leq n} \angle\big(T_{Y_j}M, \hat{T}_j\big) \leq C_{d,k,\tau_{min},\mathbf{L}} \sqrt{\frac{f_{max}}{f_{min}}} (h^{k-1} \vee \sigma h^{-1})(1 + th).$$

As a consequence, taking $t = h^{-1}$, for n large enough,

$$\sup_{P \in \mathcal{P}^k(\sigma)} \mathbb{E}_{P^{\otimes n}} \max_{1 \leq j \leq n} \angle\big(T_{Y_j}M, \hat{T}_j\big) \leq C \left(\frac{\log n}{n-1}\right)^{\frac{k-1}{d}} \left\{1 \vee \sigma \left(\frac{\log n}{n-1}\right)^{-\frac{k}{d}}\right\},$$

where $C = C_{d,k,\tau_{min},\mathbf{L},f_{min},f_{max}}$.

The proof of Theorem 2 is given in Section 5.1.2. The same bound holds for the estimation of $T_y M$ at a prescribed $y \in M$ in the model $\mathcal{P}^k_{(y)}(\sigma)$. For that, simply take $P_n^{(y)} = 1/n \sum_i \delta_{(X_i - y)}$ as integration in (2).

In the noise-free setting, or when $\sigma \leq h^k$, this result is in line with those of [6] in terms of the sample size dependency $(1/n)^{(k-1)/d}$. Besides, it shows that the convergence rate of our estimator does not depend on the ambient dimension D, even in codimension greater than 2. When $k = 2$, we recover the same rate as [1], where we used local PCA, which is a reformulation of (2). When $k \geq 3$, the procedure (2) outperforms PCA-based estimators of [25] and [27], where convergence rates of the form $(1/n)^{\beta}$ with $1/d < \beta < 2/d$ are obtained. This bound also recovers the result of [7] in the case $k = 3$, where a similar procedure is used. When the noise level σ is of order h^{α}, with $1 \leq \alpha \leq k$, Theorem 2 yields a convergence rate in $h^{\alpha-1}$. Since a polynomial decomposition up to order $k_\alpha = \lceil \alpha \rceil$ in (2) results in the same bound, the noise level $\sigma = h^{\alpha}$ may be thought of as an α-regularity threshold. At last, it may be worth mentioning that the results of Theorem 2 also hold when the assumption $\mathbb{E}(Z|Y) = 0$ is relaxed. Theorem 2 nearly matches the following lower bound.

Theorem 3. *If $\tau_{min} L_\perp, \ldots, \tau_{min}^{k-1} L_k, (\tau_{min}^d f_{min})^{-1}$ and $\tau_{min}^d f_{max}$ are large enough (depending only on d and k), then*

$$\inf_{\hat{T}} \sup_{P \in \mathcal{P}^k(\sigma)} \mathbb{E}_{P^{\otimes n}} \angle\big(T_{\pi_M(X_1)} M, \hat{T}\big) \geq c_{d,k,\tau_{min}} \left\{ \left(\frac{1}{n-1}\right)^{\frac{k-1}{d}} \vee \left(\frac{\sigma}{n-1}\right)^{\frac{k-1}{d+k}} \right\},$$

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

A proof of Theorem 3 can be found in Section 5.2.2. When $\sigma \lesssim (1/n)^{k/d}$, the lower bound matches Theorem 2 in the noise-free case, up to a $\log n$ factor. Thus, the rate $(1/n)^{(k-1)/d}$ is optimal for tangent space estimation on the model $\mathcal{P}^k$. The rate $(\log n/n)^{1/d}$ obtained in [1] for $k = 2$ is therefore optimal, as well as the rate $(\log n/n)^{2/d}$ given in [7] for $k = 3$. The rate $(1/n)^{(k-1)/d}$ naturally appears on the the model $\mathcal{P}^k$, as the estimation rate of differential objects of order 1 from k-smooth submanifolds.

When $\sigma \asymp (1/n)^{\alpha/d}$ with $\alpha < k$, the lower bound provided by Theorem 3 is of order $(1/n)^{(k-1)(\alpha+d)/[d(d+k)]}$, hence smaller than the $(1/n)^{\alpha/d}$ rate of Theorem 2. This suggests that the local polynomial estimator (2) is suboptimal whenever $\sigma \gg (1/n)^{k/d}$ on the model $\mathcal{P}^k(\sigma)$.

Here again, the same lower bound holds for the estimation of T_yM at a fixed point y in the model $\mathcal{P}^k_{(y)}(\sigma)$.

3.2. *Curvature.* The second fundamental form $II^M_{Y_j} : T_{Y_j}M \times T_{Y_j}M \to T_{Y_j}M^\perp \subset \mathbb{R}^D$ is a symmetric bilinear map that encodes completely the curvature of M at Y_j [9, Chap. 6, Proposition 3.1]. Estimating it only from a point cloud $\mathbb{X}_n$ does not trivially make sense, since $II^M_{Y_j}$ has domain $T_{Y_j}M$ which is unknown. To bypass this issue we extend $II^M_{Y_j}$ to $\mathbb{R}^D$. That is, we consider the estimation of $II^M_{Y_j} \circ \pi_{T_{Y_j}M}$ which has full domain $\mathbb{R}^D$. Following the same ideas as in the previous Section 3.1, we use the second order tensor $\hat{T}_{2,j} \circ \hat{\pi}_j$ obtained in (2) to estimate $II^M_{Y_j} \circ \pi_{T_{Y_j}M}$.

THEOREM 4. *Let $k \geq 3$. Take h as in Theorem 2, $\sigma \leq h/4$, and $t = 1/h$. If n is large enough so that $h \leq h_0$ and $h^{-1} \geq C^{-1}_{k,d,\tau_{min},\mathbf{L}} \geq (\sup_{2\leq i\leq k} \|T_i^*\|_{op})^{-1}$, then with probability at least $1 - \left(\frac{1}{n}\right)^{k/d}$,*

$$\max_{1\leq j\leq n} \left\| II^M_{Y_j} \circ \pi_{T_{Y_j}M} - \hat{T}_{2,j} \circ \hat{\pi}_j \right\|_{op} \leq C_{d,k,\tau_{min},\mathbf{L}} \sqrt{\frac{f_{max}}{f_{min}}} (h^{k-2} \vee \sigma h^{-2}).$$

In particular, for n large enough,

$$\sup_{P\in\mathcal{P}^k(\sigma)} \mathbb{E}_{P^{\otimes n}} \max_{1\leq j\leq n} \left\| II^M_{Y_j} \circ \pi_{T_{Y_j}M} - \hat{T}_{2,j} \circ \hat{\pi}_j \right\|_{op}$$
$$\leq C_{d,k,\tau_{min},\mathbf{L},f_{min},f_{max}} \left(\frac{\log n}{n-1}\right)^{\frac{k-2}{d}} \left\{ 1 \vee \sigma \left(\frac{\log n}{n-1}\right)^{-\frac{k}{d}} \right\}.$$

The proof of Theorem 4 is given in Section 5.1.3. As in Theorem 2, the case $\sigma \leq h^k$ may be thought of as a noise-free setting, and provides an upper bound of the form h^{k-2}. Interestingly, Theorems 2 and 4 are enough to provide estimators of various notions of curvature. For instance, consider the scalar curvature [9, Section 4.4] at a point Y_j, defined by

$$Sc^M_{Y_j} = \frac{1}{d(d-1)} \sum_{r\neq s} \left[\left\langle II^M_{Y_j}(e_r, e_r), II^M_{Y_j}(e_s, e_s) \right\rangle - \|II^M_{Y_j}(e_r, e_s)\|^2 \right],$$

where $(e_r)_{1\leq r\leq d}$ is an orthonormal basis of $T_{Y_j}M$. A plugin estimator of $Sc^M_{Y_j}$ is

$$\widehat{Sc}_j = \frac{1}{d(d-1)} \sum_{r\neq s} \left[\left\langle \hat{T}_{2,j}(\hat{e}_r, \hat{e}_r), \hat{T}_{2,j}(\hat{e}_s, \hat{e}_s) \right\rangle - \|\hat{T}_{2,j}(\hat{e}_r, \hat{e}_s)\|^2 \right],$$

where $(\hat{e}_r)_{1\leq r\leq d}$ is an orthonormal basis of $\hat{T}_{2,j}$. Theorems 2 and 4 yield

$$\mathbb{E}_{P^{\otimes n}} \max_{1\leq j\leq n} \left|\widehat{Sc}_j - Sc^M_{Y_j}\right| \leq C \left(\frac{\log n}{n-1}\right)^{\frac{k-2}{d}} \left\{1 \vee \sigma \left(\frac{\log n}{n-1}\right)^{-\frac{k}{d}}\right\},$$

where $C = C_{d,k,\tau_{min},\mathbf{L},f_{min},f_{max}}$.

The (near-)optimality of the bound stated in Theorem 4 is assessed by the following lower bound.

THEOREM 5. *If $\tau_{min}L_\perp, \ldots, \tau_{min}^{k-1}L_k, (\tau_{min}^d f_{min})^{-1}$ and $\tau_{min}^d f_{max}$ are large enough (depending only on d and k), then*

$$\inf_{\widehat{II}} \sup_{P\in\mathcal{P}^k(\sigma)} \mathbb{E}_{P^{\otimes n}} \left\| II^M_{\pi_M(X_1)} \circ \pi_{T_{\pi_M(X_1)}M} - \widehat{II}\right\|_{op}$$
$$\geq c_{d,k,\tau_{min}} \left\{\left(\frac{1}{n-1}\right)^{\frac{k-2}{d}} \vee \left(\frac{\sigma}{n-1}\right)^{\frac{k-2}{d+k}}\right\},$$

where the infimum is taken over all the estimators $\widehat{II} = \widehat{II}(X_1, \ldots, X_n)$.

The proof of Theorem 5 is given in Section 5.2.2. The same remarks as in Section 3.1 hold. If the estimation problem consists in approximating II^M_y at a fixed point y known to belong to M beforehand, we obtain the same rates. The ambient dimension D still plays no role. The shift $k-2$ in the rate of convergence on a $\mathcal{C}^k$-model can be interpreted as the order of derivation of the object of interest, that is 2 for curvature.

Notice that the lower bound (Theorem 5) does not require $k \geq 3$. Hence, we get that for $k = 2$, curvature cannot be estimated uniformly consistently on the $\mathcal{C}^2$-model $\mathcal{P}^2$. This seems natural, since the estimation of a second order quantity should require an additional degree of smoothness.

3.3. *Support Estimation.* For each $1 \leq j \leq n$, the minimization (2) outputs a series of tensors $(\hat{\pi}_j, \hat{T}_{2,j}, \ldots, \hat{T}_{k-1,j})$. This collection of multidimensional monomials can be further exploited as follows. By construction, they fit M at scale h around Y_j, so that

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

is a good candidate for an approximate parametrization in a neighborhood of Y_j. We do not know the domain $T_{Y_j}M$ of the initial parametrization,

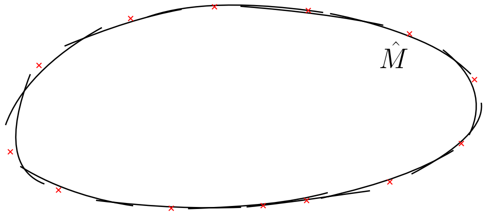


Figure 4: $\hat{M}$ is a union of polynomial patches at sample points.

though we have at hand an approximation $\hat{T}_j = \operatorname{im} \hat{\pi}_j$ which was proved to be consistent in Section 3.1. As a consequence, we let the support estimator based on local polynomials $\hat{M}$ be

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

The set $\hat{M}$ has no reason to be globally smooth, since it consists of a mere union of polynomial patches (Figure 4). However, $\hat{M}$ is provably close to M for the Hausdorff distance.

Theorem 6. *With the same assumptions as Theorem 4, with probability at least* $1 - 2\left(\frac{1}{n}\right)^{\frac{k}{d}}$, *we have*

$$d_H\big(M, \hat{M}\big) \leq C_{d,k,\tau_{min},\mathbf{L},f_{min},f_{max}}(h^k \vee \sigma).$$

In particular, for n large enough,

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

A proof of Theorem 6 is given in Section 5.1.4. As in Theorem 2, for a noise level of order h^α, $\alpha \geq 1$, Theorem 6 yields a convergence rate of order $h^{(k \wedge \alpha)/d}$. Thus the noise level σ may also be thought of as a regularity threshold. Contrary to [15, Theorem 2], the case $h/4 < \sigma < \tau_{min}$ is not in the scope of Theorem 6. Moreover, for $1 \leq \alpha < 2d/(d+2)$, [15, Theorem 2] provides a better convergence rate of $h^{2/(d+2)}$. Note however that Theorem 6 is also valid whenever the assumption $\mathbb{E}(Z|Y) = 0$ is relaxed. In this non-centered noise framework, Theorem 6 outperforms [20, Theorem 7] in the case $d \geq 3$, $k = 2$, and $\sigma \leq h^2$.

In the noise-free case or when $\sigma \leq h^k$, for $k = 2$, we recover the rate $(\log n/n)^{2/d}$ obtained in [1, 14, 19] and improve the rate $(\log n/n)^{2/(d+2)}$ in

[15, 20]. However, our estimator $\hat{M}$ is an unstructured union of d-dimensional balls in $\mathbb{R}^D$. Consequently, $\hat{M}$ does not recover the topology of M as the estimator of [1] does.

When $k \geq 3$, $\hat{M}$ outperforms reconstruction procedures based on a somewhat piecewise linear interpolation [1, 14, 20], and achieves the faster rate $(\log n/n)^{k/d}$ for the Hausdorff loss. This seems quite natural, since our procedure fits higher order terms. This is done at the price of a probably worse dependency on the dimension d than in [1, 14]. Theorem 6 is now proved to be (almost) minimax optimal.

Theorem 7. *If $\tau_{min}L_{\perp}, \ldots, \tau_{min}^{k-1}L_k, (\tau_{min}^d f_{min})^{-1}$ and $\tau_{min}^d f_{max}$ are large enough (depending only on d and k), then for n large enough,*

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}^k(\sigma)} \mathbb{E}_{P^{\otimes n}} d_H\big(M, \hat{M}\big) \geq c_{d,k,\tau_{min}} \left\{ \left(\frac{1}{n}\right)^{\frac{k}{d}} \vee \left(\frac{\sigma}{n}\right)^{\frac{k}{d+k}} \right\},$$

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

Theorem 7, whose proof is given in Section 5.2.1, is obtained from Le Cam's Lemma (Theorem 8). Let us note that it is likely for the extra $\log n$ term appearing in Theorem 6 to actually be present in the minimax rate. Roughly, it is due to the fact that the Hausdorff distance d_H is similar to a L^∞ loss. The $\log n$ term may be obtained in Theorem 7 with the same combinatorial analysis as in [19] for $k = 2$.

As for the estimation of tangent spaces and curvature, Theorem 7 matches the upper bound in Theorem 6 in the noise-free case $\sigma \lesssim (1/n)^{k/d}$. Moreover, for $\sigma < \tau_{min}$, it also generalizes Theorem 1 in [15] to higher orders of regularity ($k \geq 3$). Again, for $\sigma \gg (1/n)^{-k/d}$, the upper bound in Theorem 6 is larger than the lower bound stated in Theorem 7. However our estimator $\hat{M}$ achieves the same convergence rate if the assumption $\mathbb{E}(Z|Y)$ is dropped.

4. Conclusion, Prospects.

In this article, we derived non-asymptotic bounds for inference of geometric objects associated with smooth submanifolds $M \subset \mathbb{R}^D$. We focused on tangent spaces, second fundamental forms, and the submanifold itself. We introduced new regularity classes $\mathcal{C}^k_{\tau_{min},\mathbf{L}}$ for submanifolds that extend the case $k = 2$. For each object of interest, the proposed estimator relies on local polynomials that can be computed through a least square minimization. Minimax lower bounds were presented, matching the upper bounds up to $\log n$ factors in the regime of small noise.

The implementation of (2) needs to be investigated. The non-convexity of the criterion comes from that we minimize over the space of orthogonal projectors, which is non-convex. However, that space is pretty well understood, and it seems possible to implement gradient descents on it [28]. Another way to improve our procedure could be to fit orthogonal polynomials instead of monomials. Such a modification may also lead to improved dependency on the dimension d and the regularity k in the bounds for both tangent space and support estimation.

Though the stated lower bounds are valid for quite general tubular noise levels σ, it seems that our estimators based on local polynomials are suboptimal whenever σ is larger than the expected precision for $\mathcal{C}^k$ models in a d-dimensional space (roughly $(1/n)^{k/d}$). In such a setting, it is likely that a preliminary centering procedure is needed, as the one exposed in [15]. Other pre-processings of the data might adapt our estimators to other types of noise. For instance, whenenever outliers are allowed in the model $\mathcal{C}^2$, [1] proposes an iterative denoising procedure based on tangent space estimation. It exploits the fact that tangent space estimation allows to remove a part of outliers, and removing outliers enhances tangent space estimation. An interesting question would be to study how this method can apply with local polynomials.

Another open question is that of exact topology recovery with fast rates for $k \geq 3$. Indeed, $\hat{M}$ converges at rate $(\log n/n)^{k/d}$ but is unstructured. It would be nice to glue the patches of $\hat{M}$ together, for example using interpolation techniques, following the ideas of [12].

5. Proofs.

5.1. *Upper bounds.*

5.1.1. *Preliminary results on polynomial expansions.* To prove Theorem 2, 4 and 6, the following lemmas are needed. First, we relate the existence of parametrizations Ψ_p's mentioned in Definition 1 to a local polynomial decomposition.

Lemma 2. *For any $M \in \mathcal{C}^k_{\tau_{min},\mathbf{L}}$ and $y \in M$, the following holds.*

(i) For all $v_1, v_2 \in \mathcal{B}_{T_yM}\left(0, \frac{1}{4L_\perp}\right)$,

$$\frac{3}{4}\|v_2 - v_1\| \leq \|\Psi_y(v_2) - \Psi_y(v_1)\| \leq \frac{5}{4}\|v_2 - v_1\|.$$

(ii) For all $h \leq \frac{1}{4L_\perp} \wedge \frac{2\tau_{min}}{5}$,

$$M \cap \mathcal{B}\left(y, \frac{3h}{5}\right) \subset \Psi_y\left(\mathcal{B}_{T_yM}(y,h)\right) \subset M \cap \mathcal{B}\left(y, \frac{5h}{4}\right).$$

(iii) For all $h \leq \frac{\tau_{min}}{2}$,

$$\mathcal{B}_{T_yM}\left(0, \frac{7h}{8}\right) \subset \pi_{T_yM}\left(\mathcal{B}(y,h) \cap M\right).$$

(iv) Denoting by $\pi^* = \pi_{T_yM}$ *the orthogonal projection onto* T_yM, *for all* $y \in M$, *there exist multilinear maps* $T_2^*, \ldots, T_{k-1}^*$ *from* T_yM *to* $\mathbb{R}^D$, *and* R_k *such that for all* $y' \in \mathcal{B}\left(y, \frac{\tau_{min} \wedge L_\perp^{-1}}{4}\right) \cap M$,

$$\begin{aligned} y' - y = \pi^*(y'-y) + T_2^*(\pi^*(y'-y)^{\otimes 2}) + \ldots + T_{k-1}^*(\pi^*(y'-y)^{\otimes k-1}) \\ + R_k(y'-y), \end{aligned}$$

with

$$\left\|R_k(y'-y)\right\| \leq C\left\|y'-y\right\|^k \qquad \text{and} \qquad \|T_i^*\|_{op} \leq L_i', \text{ for } 2 \leq i \leq k-1,$$

where L_i' *depends on* $d, k, \tau_{min}, L_\perp, \ldots, L_i$, *and* C *on* d, k, τ_{min}, $L_\perp, \ldots$, L_k. *Moreover, for* $k \geq 3$, $T_2^* = II_y^M$.

(v) For all $y \in M$, $\left\|II_y^M\right\|_{op} \leq 1/\tau_{min}$. *In particular, the sectional curvatures of* M *satisfy*

$$\frac{-2}{\tau_{min}^2} \leq \kappa \leq \frac{1}{\tau_{min}^2}.$$

The proof of Lemma 2 can be found in Section A.2 of the Appendix. A direct consequence of Lemma 2 is the following Lemma 3.

LEMMA 3. *Set* $h_0 = (\tau_{min} \wedge L_\perp^{-1})/8$ *and* $h \leq h_0$. *Let* $M \in \mathcal{C}^k_{\tau_{min},\mathbf{L}}$, $x_0 = y_0 + z_0$, *with* $y_0 \in M$ *and* $\|z_0\| \leq \sigma \leq h/4$. *Denote by* π^* *the orthogonal projection onto* $T_{y_0}M$, *and by* $T_2^*, \ldots, T_{k-1}^*$ *the multilinear maps given by Lemma 2, iv).*

Then, for any $x = y + z$ *such that* $y \in M$, $\|z\| \leq \sigma \leq h/4$ *and* $x \in \mathcal{B}(x_0, h)$, *for any orthogonal projection* π *and multilinear maps* $T_2, \ldots, T_{k-1}$, *we have*

$$\begin{aligned} x - x_0 - \pi(x-x_0) - \sum_{j=2}^{k-1} T_j(\pi(x-x_0)^{\otimes j}) = \sum_{j=1}^{k} T_j'(\pi^*(y-y_0)^{\otimes j}) \\ + R_k(x-x_0), \end{aligned}$$

where T_j' are j-linear maps, and $\|R_k(x-x_0)\| \leq C\left(\sigma \vee h^k\right)(1+th)$, with $t = \max_{j=2\dots,k}\|T\|_{op}$ and C depending on d, k, τ_{min}, $L_\perp$,…, L_k. Moreover, we have

$$\begin{array}{rcc} T_1' & = & (\pi^* - \pi), \\ T_2' & = & (\pi^*-\pi)\circ T_2^* + (T_2^*\circ\pi^* - T_2\circ\pi), \end{array}$$

and, if $\pi = \pi^*$ and $T_i = T_i^*$, for $i = 2,\dots,k-1$, then $T_j' = 0$, for $j = 1,\dots,k$.

Lemma 3 roughly states that, if π, T_j, $j \geq 2$ are designed to locally approximate $x = y + z$ around $x_0 = y_0 + z_0$, then the approximation error may be expressed as a polynomial expansion in $\pi^*(y - y_0)$.

Proof of Lemma 3. For short assume that $y_0 = 0$. In what follows C will denote a constant depending on d, k, τ_{min}, $L_\perp$,…, L_k. We may write

$$x - x_0 - \pi(x-x_0) - \sum_{j=2}^{k-1} T_j(\pi(x-x_0)^{\otimes j}) = y - \pi(y) - \sum_{j=2}^{k-1} T_j(\pi(y)^{\otimes j}) \\ + R_k'(x-x_0),$$

with $\|R_k'(x-x_0)\| \leq C\sigma(1+th)$. Since $\sigma \leq h/4$, $y \in \mathcal{B}(0, 3h/2)$, with $h \leq h_0$. Hence Lemma 2 entails

$$y = \pi^*(y) + T_2^*(\pi^*(y)^{\otimes 2}) + \ldots + T_{k-1}^*(\pi^*(y)^{\otimes k-1}) \\ + R_k''(y),$$

with $\|R_k''(y)\| \leq Ch^k$. We deduce that

$$y - \pi(y) - \sum_{j=2}^{k-1} T_j(\pi(y)^{\otimes j}) = (\pi^* - \pi\circ\pi^*)(y) + T_2^*(\pi^*(y)^{\otimes 2}) - \pi(T_2^*(\pi^*(y)^{\otimes 2})) \\ - T_2(\pi\circ\pi^*(y)^{\otimes 2}) + \sum_{j=3}^{k} T_k'(\pi^*(y)^{\otimes j}) - \pi(R_k''(y)) - R_k'''(y),$$

with $\|R_k'''(y)\| \leq Cth^{k+1}$, since only tensors of order greater than 2 are involved in R_k'''. Since $T_2^* = II_0^M$, $\pi^*\circ T_2^* = 0$, hence the result. □

At last, we need a result relating deviation in terms of polynomial norm and $L^2(P_{0,n-1}^{(j)})$ norm, where $P_0 \in \mathcal{P}^k$, for polynomials taking arguments in $\pi^{*,(j)}(y)$. For clarity's sake, the bounds are given for $j = 1$, and we denote $P_{0,n-1}^{(1)}$ by $P_{0,n-1}$. Without loss of generality, we can assume that $Y_1 = 0$.

Let $\mathbb{R}^k[y_{1:d}]$ denote the set of real-valued polynomial functions in d variables with degree less than k. For $S \in \mathbb{R}^k[y_{1:d}]$, we denote by $\|S\|_2$ the Euclidean norm of its coefficients, and by S_h the polynomial defined by $S_h(y_{1:d}) = S(hy_{1:d})$. With a slight abuse of notation, $S(\pi^*(y))$ will denote $S(e_1^*(\pi^*(y)), \ldots, e_d^*(\pi^*(y)))$, where $e_1^*, \ldots, e_d^*$ form an orthonormal coordinate system of $T_0 M$.

Proposition 2. *Set* $h = \left(K \frac{\log n}{n-1}\right)^{\frac{1}{d}}$. *There exist constants* $\kappa_{k,d}$, $c_{k,d}$ *and* C_d *such that, if* $K \geq (\kappa_{k,d} f_{max}^2 / f_{min}^3)$ *and* n *is large enough so that* $h \leq h_0 \leq \tau_{min}/8$, *then with probability at least* $1 - \left(\frac{1}{n}\right)^{\frac{k}{d}+1}$, *we have*

$$\begin{array}{rcl} P_{0,n-1}[S^2(\pi^*(y))\mathbb{1}_{\mathcal{B}(h/2)}(y)] & \geq & c_{k,d} h^d f_{min} \|S_h\|_2^2, \\ N(3h/2) & \leq & C_d f_{max} (n-1) h^d, \end{array}$$

for every $S \in \mathbb{R}^k[y_{1:d}]$, *where* $N(3h/2) = \sum_{j=2}^n \mathbb{1}_{\mathcal{B}(0,3h/2)}(Y_j)$.

The proof of Proposition 2 is deferred to Section B.2 of the Appendix.

5.1.2. *Upper Bound for Tangent Space Estimation.*

Proof of Theorem 2. We recall that for every $j = 1, \ldots, n$, $X_j = Y_j + Z_j$, where $Y_j \in M$ is drawn from P_0 and $\|Z_j\| \leq \sigma \leq h/4$, where $h \leq h_0$ as defined in Lemma 3. Without loss of generality we consider the case $j = 1$, $Y_1 = 0$. From now on we assume that the probability event defined in Proposition 2 occurs, and denote by $\mathcal{R}_{n-1}(\pi, T_2, \ldots, T_{k-1})$ the empirical criterion defined by (2). Note that $X_j \in \mathcal{B}(X_1, h)$ entails $Y_j \in \mathcal{B}(0, 3h/2)$. Moreover, since for $t \geq \max_{i=2,\ldots,k-1} \|T_i^*\|_{op}$, $\mathcal{R}_{n-1}(\hat{\pi}, \hat{T}_1, \ldots, \hat{T}_{k-1}) \leq \mathcal{R}_{n-1}(\pi^*, T_2^*, \ldots, T_{k-1}^*)$, we deduce that

$$\mathcal{R}_{n-1}(\hat{\pi}, \hat{T}_1, \ldots, \hat{T}_{k-1}) \leq \frac{C_{\tau_{min}, \mathbf{L}} \left(\sigma^2 \vee h^{2k}\right)(1+th)^2 N(3h/2)}{n-1},$$

according to Lemma 3. On the other hand, note that if $Y_j \in \mathcal{B}(0, h/2)$, then $X_j \in \mathcal{B}(X_1, h)$. Lemma 3 then yields

$$\begin{aligned} \mathcal{R}_{n-1}(\hat{\pi}, \hat{T}_2, \ldots, \hat{T}_{k-1}) \geq & P_{0,n-1}\left(\left\| \sum_{j=1}^k \hat{T}_j'(\pi^*(y)^{\otimes j}) \right\|^2 \mathbb{1}_{\mathcal{B}(0,h/2)}(y) \right) \\ & - \frac{C_{\tau_{min}, \mathbf{L}} \left(\sigma^2 \vee h^{2k}\right)(1+th)^2 N(3h/2)}{n-1}. \end{aligned}$$

Using Proposition 2, we can decompose the right-hand side as

$$\sum_{r=1}^{D} P_{0,n-1}\left(\sum_{j=1}^{k}\hat{T}'^{(r)}_{j}(\pi^*(y)^{\otimes j})\mathbb{1}_{\mathcal{B}(0,h/2)}(y)\right)^2 \leq C_{\tau_{min},\mathbf{L}} f_{max} h^d \left(\sigma^2 \vee h^{2k}\right)(1+th)^2,$$

where for any tensor T, $T^{(r)}$ denotes the r-th coordinate of T and is considered as a real valued r-order polynomial. Then, applying Proposition 2 to each coordinate leads to

$$c_{d,k} f_{min} \sum_{r=1}^{D}\sum_{j=1}^{k}\left\|\left(T'^{(r)}_{j}(\pi^*(y)^{\otimes j})\right)_h\right\|_2^2 \leq C_{\tau_{min},\mathbf{L}} f_{max} h^d \left(\sigma^2 \vee h^{2k}\right)(1+th)^2.$$

It follows that, for $1 \leq j \leq k$,

$$\|\hat{T}'_j \circ \pi^*\|_{op}^2 \leq C_{d,k,\mathbf{L},\tau_{min}} \frac{f_{max}}{f_{min}} (h^{2(k-j)} \vee \sigma^2 h^{-2j})(1+t^2h^2). \tag{3}$$

Noting that, according to [16, Section 2.6.2],

$$\|\hat{T}'_1 \circ \pi^*\|_{op} = \|(\pi^* - \hat{\pi})\pi^*\|_{op} = \|\pi_{\hat{T}_1^{\perp}} \circ \pi^*\| = \angle(T_{Y_1}M, \hat{T}_1),$$

we deduce that

$$\angle\big(T_{Y_1}M, \hat{T}_1\big) \leq C_{d,k,\mathbf{L},\tau_{min}} \sqrt{\frac{f_{max}}{f_{min}}} (h^{(k-1)} \vee \sigma h^{-1})(1+th).$$

Theorem 2 then follows from a straightforward union bound. □

5.1.3. *Upper Bound for Curvature Estimation.*

Proof of Theorem 4. Without loss of generality, the derivation is conducted in the same framework as in the previous Section 5.1.2. In accordance with assumptions of Theorem 4, we assume that $\max_{2 \leq i \leq k} \|T_i^*\|_{op} \leq t \leq 1/h$. Since, according to Lemma 3,

$$T_2'(\pi^*(y)^{\otimes 2}) = (\pi^* - \hat{\pi})(T_2^*(\pi^*(y)^{\otimes 2})) + (T_2^* \circ \pi^* - \hat{T}_2 \circ \hat{\pi})(\pi^*(y)^{\otimes 2}),$$

we deduce that

$$\|T_2^* \circ \pi^* - \hat{T}_2 \circ \hat{\pi}\|_{op} \leq \|T_2' \circ \pi^*\|_{op} + \|\hat{\pi} - \pi^*\|_{op} + \|\hat{T}_2 \circ \hat{\pi} \circ \pi^* - \hat{T}_2 \circ \hat{\pi} \circ \hat{\pi}\|_{op}.$$

Using (3) with $j = 1, 2$ and $th \leq 1$ leads to

$$\|T_2^* \circ \pi^* - \hat{T}_2 \circ \hat{\pi}\|_{op} \leq C_{d,k,\mathbf{L},\tau_{min}} \sqrt{\frac{f_{max}}{f_{min}}} (h^{(k-2)} \vee \sigma h^{-2}).$$

Finally, Lemma 2 states that $II_{Y_1}^M = T_2^*$. Theorem 4 follows from a union bound. □

5.1.4. *Upper Bound for Manifold Estimation.*

Proof of Theorem 6 . Recall that we take $X_i = Y_i + Z_i$, where Y_i has distribution P_0 and $\|Z_j\| \leq \sigma \leq h/4$. We also assume that the probability events of Proposition 2 occur simultaneously at each Y_i, so that (3) holds for all i, with probability larger than $1-(1/n)^{k/d}$. Without loss of generality set $Y_1 = 0$. Let $v \in \mathcal{B}_{\hat{T}_1 M}(0, 7h/8)$ be fixed. Notice that $\pi^*(v) \in \mathcal{B}_{T_0 M}(0, 7h/8)$. Hence, according to Lemma 2, there exists $y \in \mathcal{B}(0, h) \cap M$ such that $\pi^*(v) = \pi^*(y)$. According to (3), we may write

$$\widehat{\Psi}(v) = Z_1 + v + \sum_{j=2}^{k-1} \hat{T}_j(v^{\otimes j}) = \pi^*(v) + \sum_{j=2}^{k-1} \hat{T}_j(\pi^*(v)^{\otimes j}) + R_k(v),$$

where, since $\|\hat{T}_j\|_{op} \leq 1/h$, $\|R_k(v)\| \leq C_{k,d,\tau_{min},\mathbf{L}} \sqrt{f_{max}/f_{min}}(h^k \vee \sigma)$. Using (3) again leads to

$$\begin{aligned} \pi^*(v) + \sum_{j=2}^{k-1} \hat{T}_j(\pi^*(v)^{\otimes j}) &= \pi^*(v) + \sum_{j=2}^{k-1} T_j^*(\pi^*(v)^{\otimes j}) + R'(\pi^*(v)) \\ &= \pi^*(y) + \sum_{j=2}^{k-1} T_j^*(\pi^*(y)^{\otimes j}) + R'(\pi^*(y)), \end{aligned}$$

where $\|R'(\pi^*(y))\| \leq C_{k,d,\tau_{min},\mathbf{L}} \sqrt{f_{max}/f_{min}}(h^k \vee \sigma)$. According to Lemma 2, we deduce that $\|\widehat{\Psi}(v) - y\| \leq C_{k,d,\tau_{min},\mathbf{L}} \sqrt{f_{max}/f_{min}}(h^k \vee \sigma)$, hence

$$\sup_{u \in \hat{M}} d(u, M) \leq C_{k,d,\tau_{min},\mathbf{L}} \sqrt{\frac{f_{max}}{f_{min}}} (h^k \vee \sigma). \tag{4}$$

Now we focus on $\sup_{y \in M} d(y, \hat{M})$. For this, we need a lemma ensuring that $\mathbb{Y}_n = \{Y_1, \ldots, Y_n\}$ covers M with high probability.

LEMMA 4. *Let $h = \left(\frac{C'_d k}{f_{min}} \frac{\log n}{n}\right)^{1/d}$ with C'_d large enough. Then for n large enough so that $h \leq \tau_{min}/4$, with probability at least $1 - \left(\frac{1}{n}\right)^{k/d}$,*

$$d_H\left(M, \mathbb{Y}_n\right) \leq h/2.$$

The proof of Lemma 4 is given in Section B.1 of the Appendix. Now we choose h satisfying the conditions of Proposition 2 and Lemma 4. Let y be in M and assume that $\|y - Y_{j_0}\| \leq h/2$. Then $y \in \mathcal{B}(X_{j_0}, 3h/4)$. According to Lemma 3 and (3), we deduce that $\|\widehat{\Psi}_{j_0}(\hat{\pi}_{j_0}(y - X_{j_0})) - y\| \leq C_{k,d,\tau_{min},\mathbf{L}}\sqrt{f_{max}/f_{min}}(h^k \vee \sigma)$. Hence, from Lemma 4,

$$\sup_{y \in M} d(y, \hat{M}) \leq C_{k,d,\tau_M,\mathbf{L}}\sqrt{\frac{f_{max}}{f_{min}}}(h^k \vee \sigma) \tag{5}$$

with probability at least $1 - 2\left(\frac{1}{n}\right)^{k/d}$. Combining (4) and (5) gives Theorem 6. □

5.2. *Minimax Lower Bounds.* This section is devoted to describe the main ideas of the proofs of the minimax lower bounds. We prove Theorem 7 on one side, and Theorem 3 and Theorem 5 in a unified way on the other side. The methods used rely on hypothesis comparison [30].

5.2.1. *Lower Bound for Manifold Estimation.* We recall that for two distributions Q and Q' defined on the same space, the L^1 test affinity $\|Q \wedge Q'\|_1$ is given by

$$\left\|Q \wedge Q'\right\|_1 = \int dQ \wedge dQ',$$

where dQ and dQ' denote densities of Q and Q' with respect to any dominating measure.

The first technique we use, involving only two hypotheses, is usually referred to as Le Cam's Lemma [30]. Let $\mathcal{P}$ be a model and $\theta(P)$ be the parameter of interest. Assume that $\theta(P)$ belongs to a pseudo-metric space $(\mathcal{D}, d)$, that is $d(\cdot,\cdot)$ is symmetric and satisfies the triangle inequality. Le Cam's Lemma can be adapted to our framework as follows.

THEOREM 8 (Le Cam's Lemma [30]). *For all pairs P, P' in $\mathcal{P}$,*

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{\otimes n}} d(\theta(P), \hat{\theta}) \geq \frac{1}{2} d\left(\theta(P), \theta(P')\right) \left\|P \wedge P'\right\|_1^n,$$

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

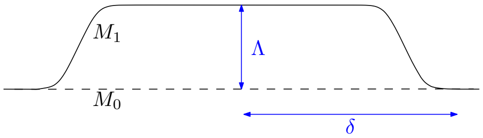


Figure 5: Manifolds M_0 and M_1 of Lemma 5 and Lemma 6. The width δ of the bump is chosen to have $\|P_0^\sigma \wedge P_1^\sigma\|_1^n$ constant. The distance $\Lambda = d_H(M_0, M_1)$ is of order δ^k to ensure that $M_1 \in \mathcal{C}^k$.

In this section, we will get interested in $\mathcal{P} = \mathcal{P}^k(\sigma)$ and $\theta(P) = M$, with $d = d_H$. In order to derive Theorem 7, we build two different pairs (P_0, P_1), (P_0^σ, P_1^σ) of hypotheses in the model $\mathcal{P}^k(\sigma)$. Each pair will exploit a different property of the model $\mathcal{P}^k(\sigma)$.

The first pair (P_0, P_1) of hypotheses (Lemma 5) is built in the model $\mathcal{P}^k \subset \mathcal{P}^k(\sigma)$, and exploits the geometric difficulty of manifold reconstruction, even if no noise is present. These hypotheses, depicted in Figure 5, consist of bumped versions of one another.

Lemma 5. *Under the assumptions of Theorem 7, there exist $P_0, P_1 \in \mathcal{P}^k$ with associated submanifolds M_0, M_1 such that*

$$d_H(M_0, M_1) \geq c_{k,d,\tau_{min}} \left(\frac{1}{n}\right)^{\frac{k}{d}}, \text{ and } \quad \|P_0 \wedge P_1\|_1^n \geq c_0.$$

The proof of Lemma 5 is to be found in Section C.4.1 of the Appendix.

The second pair (P_0^σ, P_1^σ) of hypotheses (Lemma 6) has a similar construction than (P_0, P_1). Roughly speaking, they are the uniform distributions on the offsets of radii $\sigma/2$ of M_0 and M_1 of Figure 5. Here, the hypotheses are built in $\mathcal{P}^k(\sigma)$, and fully exploit the statistical difficulty of manifold reconstruction induced by noise.

Lemma 6. *Under the assumptions of Theorem 7, there exist $P_0^\sigma, P_1^\sigma \in \mathcal{P}^k(\sigma)$ with associated submanifolds M_0^σ, M_1^σ such that*

$$d_H(M_0^\sigma, M_1^\sigma) \geq c_{k,d,\tau_{min}} \left(\frac{\sigma}{n}\right)^{\frac{k}{d+k}}, \text{ and } \quad \|P_0^\sigma \wedge P_1^\sigma\|_1^n \geq c_0.$$

The proof of Lemma 6 is to be found in Section C.4.2 of the Appendix. We are now in position to prove Theorem 7.

Proof of Theorem 7. Let us apply Theorem 8 with $\mathcal{P} = \mathcal{P}^k(\sigma)$, $\theta(P) = M$ and $d = d_H$. Taking $P = P_0$ and $P' = P_1$ of Lemma 5, these distributions both belong to $\mathcal{P}^k \subset \mathcal{P}^k(\sigma)$, so that Theorem 8 yields

$$\begin{aligned}\inf_{\hat{M}} \sup_{P \in \mathcal{P}^k(\sigma)} \mathbb{E}_{P^{\otimes n}} d_H\big(M, \hat{M}\big) &\geq d_H(M_0, M_1) \, \|P_0 \wedge P_1\|_1^n \\ &\geq c_{k,d,\tau_{min}} \left(\frac{1}{n}\right)^{\frac{k}{d}} \times c_0.\end{aligned}$$

Similarly, setting hypotheses $P = P_0^\sigma$ and $P' = P_1^\sigma$ of Lemma 6 yields

$$\begin{aligned}\inf_{\hat{M}} \sup_{P \in \mathcal{P}^k(\sigma)} \mathbb{E}_{P^{\otimes n}} d_H\big(M, \hat{M}\big) &\geq d_H(M_0^\sigma, M_1^\sigma) \, \|P_0^\sigma \wedge P_1^\sigma\|_1^n \\ &\geq c_{k,d,\tau_{min}} \left(\frac{\sigma}{n}\right)^{\frac{k}{k+d}} \times c_0,\end{aligned}$$

which concludes the proof. □

5.2.2. *Lower Bounds for Tangent Space and Curvature Estimation.* Let us now move to the proof of Theorem 3 and 5, that consist of lower bounds for the estimation of $T_{X_1}M$ and $II_{X_1}^M$ with random base point X_1. In both cases, the loss can be cast as

$$\begin{aligned}E_{P^{\otimes n}} \; d(\theta_{X_1}(P), \hat{\theta}) &= \mathbb{E}_{P^{\otimes n-1}} \left[E_P \; d(\theta_{X_1}(P), \hat{\theta}) \right] \\ &= \mathbb{E}_{P^{\otimes n-1}} \left[\left\| d\big(\theta_{\cdot}(P), \hat{\theta}\big) \right\|_{L^1(P)} \right],\end{aligned}$$

where $\hat{\theta} = \hat{\theta}(X, X')$, with $X = X_1$ driving the parameter of interest, and $X' = (X_2, \ldots, X_n) = X_{2:n}$. Since $\|.\|_{L^1(P)}$ obviously depends on P, the technique exposed in the previous section does not apply anymore. However, a slight adaptation of Assouad's Lemma [30] with an extra conditioning on $X = X_1$ carries out for our purpose. Let us now detail a general framework where the method applies.

We let $\mathcal{X}, \mathcal{X}'$ denote measured spaces. For a probability distribution Q on $\mathcal{X} \times \mathcal{X}'$, we let (X, X') be a random variable with distribution Q. The marginals of Q on $\mathcal{X}$ and $\mathcal{X}'$ are denoted by μ and ν respectively. Let $(\mathcal{D}, d)$ be a pseudo-metric space. For $Q \in \mathcal{Q}$, we let $\theta_{\cdot}(Q) : \mathcal{X} \to \mathcal{D}$ be defined μ-almost surely, where μ is the marginal distribution of Q on $\mathcal{X}$. The parameter of interest is $\theta_X(Q)$, and the associated minimax risk over $\mathcal{Q}$ is

$$\inf_{\hat{\theta}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \left[d\big(\theta_X(Q), \hat{\theta}(X, X')\big) \right], \tag{6}$$

where the infimum is taken over all the estimators $\hat{\theta} : \mathcal{X} \times \mathcal{X}' \to \mathcal{D}$.

Given a set of probability distributions $\mathcal{Q}$ on $\mathcal{X} \times \mathcal{X}'$, write $\overline{Conv}(\mathcal{Q})$ for the set of mixture probability distributions with components in $\mathcal{Q}$. For all $\tau = (\tau_1, \ldots, \tau_m) \in \{0,1\}^m$, τ^k denotes the m-tuple that differs from τ only at the kth position. We are now in position to state the conditional version of Assouad's Lemma that allows to lower bound the minimax risk (6).

Lemma 7 (Conditional Assouad). *Let $m \geq 1$ be an integer and let $\{\mathcal{Q}_\tau\}_{\tau \in \{0,1\}^m}$ be a family of 2^m submodels $\mathcal{Q}_\tau \subset \mathcal{Q}$. Let $\{U_k \times U_k'\}_{1 \leq k \leq m}$ be a family of pairwise disjoint subsets of $\mathcal{X} \times \mathcal{X}'$, and $\mathcal{D}_{\tau,k}$ be subsets of $\mathcal{D}$. Assume that for all $\tau \in \{0,1\}^m$ and $1 \leq k \leq m$,*

- *for all $Q_\tau \in \mathcal{Q}_\tau$, $\theta_X(Q_\tau) \in \mathcal{D}_{\tau,k}$ on the event $\{X \in U_k\}$;*
- *for all $\theta \in \mathcal{D}_{\tau,k}$ and $\theta' \in \mathcal{D}_{\tau^k,k}$, $d(\theta, \theta') \geq \Delta$.*

For all $\tau \in \{0,1\}^m$, let $\overline{Q}_\tau \in \overline{Conv}(\mathcal{Q}_\tau)$, and write $\bar{\mu}_\tau$ and $\bar{\nu}_\tau$ for the marginal distributions of $\overline{Q}_\tau$ on $\mathcal{X}$ and $\mathcal{X}'$ respectively. Assume that if (X, X') has distribution $\overline{Q}_\tau$, X and X' are independent conditionally on the event $\{(X, X') \in U_k \times U_k'\}$, and that

$$\min_{\substack{\tau \in \{0,1\}^m \\ 1 \leq k \leq m}} \left\{ \left(\int_{U_k} d\bar{\mu}_\tau \wedge d\bar{\mu}_{\tau^k} \right) \left(\int_{U_k'} d\bar{\nu}_\tau \wedge d\bar{\nu}_{\tau^k} \right) \right\} \geq 1 - \alpha.$$

Then,

$$\inf_{\hat{\theta}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \left[d\big(\theta_X(Q), \hat{\theta}(X, X')\big) \right] \geq m \frac{\Delta}{2} (1 - \alpha),$$

where the infimum is taken over all the estimators $\hat{\theta} : \mathcal{X} \times \mathcal{X}' \to \mathcal{D}$.

Note that for a model of the form $\mathcal{Q} = \{\delta_{x_0} \otimes P, P \in \mathcal{P}\}$ with fixed $x_0 \in \mathcal{X}$, one recovers the classical Assouad's Lemma [30] taking $U_k = \mathcal{X}$ and $U_k' = \mathcal{X}'$. Indeed, when $X = x$ is deterministic, the parameter of interest $\theta_X(Q) = \theta(Q)$ can be seen as non-random.

In this section, we will get interested in $\mathcal{Q} = \mathcal{P}^k(\sigma)^{\otimes n}$, and $\theta_X(Q) = \theta_{X_1}(Q)$ being alternatively $T_{X_1}M$ and $II_{X_1}^M$. Similarly to Section 5.2.1, we build two different families of submodels, each of them will exploit a different kind of difficulty for tangent space and curvature estimation.

The first family, described in Lemma 8, highlights the geometric difficulty of the estimation problems, even when the noise level σ is small, or even zero. Let us emphasize that the estimation error is integrated with respect to the distribution of X_1. Hence, considering mixture hypotheses is natural, since building manifolds with different tangent spaces (or curvature) necessarily leads to distributions that are locally singular. Here, as in Section 5.2.1, the considered hypotheses are composed of bumped manifolds (see Figure 6). We defer the proof of Lemma 8 to Section C.3.1 of the Appendix.

LEMMA 8. *Assume that the conditions of Theorem 3 or 5 hold. Given* $i \in \{1,2\}$, *there exists a family of* 2^m *submodels* $\left\{\mathcal{P}^{(i)}_\tau\right\}_{\tau \in \{0,1\}^m} \subset \mathcal{P}^k$, *together with pairwise disjoint subsets* $\{U_k \times U'_k\}_{1 \leq k \leq m}$ *of* $\mathbb{R}^D \times \left(\mathbb{R}^D\right)^{n-1}$ *such that the following holds for all* $\tau \in \{0,1\}^m$ *and* $1 \leq k \leq m$.

For any distribution $P^{(i)}_\tau \in \mathcal{P}^{(i)}_\tau$ *with support* $M^{(i)}_\tau = Supp\left(P^{(i)}_\tau\right)$, *if* $(X_1, \ldots, X_n)$ *has distribution* $\left(P^{(i)}_\tau\right)^{\otimes n}$, *then on the event* $\{X_1 \in U_k\}$, *we have:*

- *if* $\tau_k = 0$,

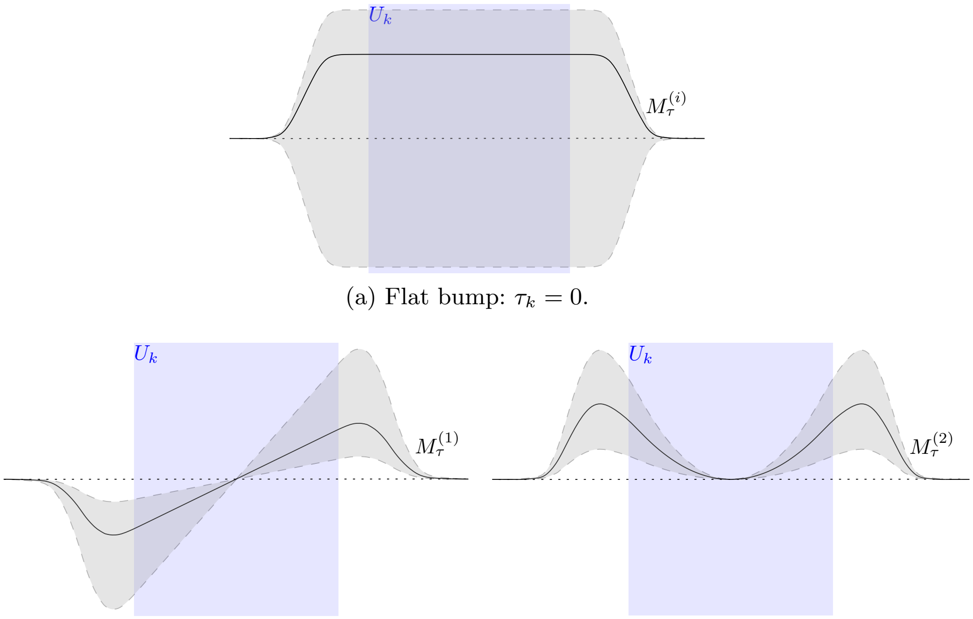
$$T_{X_1} M^{(i)}_\tau = \mathbb{R}^d \times \{0\}^{D-d} \quad , \quad \left\| II^{M^{(i)}_\tau}_{X_1} \circ \pi_{T_{X_1} M^{(i)}_\tau} \right\|_{op} = 0,$$

- *if* $\tau_k = 1$,
  - *for* $i = 1$*:* $\angle\left(T_{X_1} M^{(1)}_\tau, \mathbb{R}^d \times \{0\}^{D-d}\right) \geq c_{k,d,\tau_{min}} \left(\frac{1}{n-1}\right)^{\frac{k-1}{d}}$,
  - *for* $i = 2$*:* $\left\| II^{M^{(2)}_\tau}_{X_1} \circ \pi_{T_{X_1} M^{(2)}_\tau} \right\|_{op} \geq c_{k,d,\tau_{min}} \left(\frac{1}{n-1}\right)^{\frac{k-2}{d}}$.

Furthermore, there exists $\bar{Q}^{(i)}_{\tau,n} \in \overline{Conv}\left(\left(\mathcal{P}^{(i)}_\tau\right)^{\otimes n}\right)$ *such that if* $(Z_1, \ldots, Z_n) = (Z_1, Z_{2:n})$ *has distribution* $\bar{Q}^{(i)}_{\tau,n}$, Z_1 *and* $Z_{2:n}$ *are independent conditionally on the event* $\{(Z_1, Z_{2:n}) \in U_k \times U'_k\}$. *The marginal distributions of* $\bar{Q}^{(i)}_{\tau,n}$ *on* $\mathbb{R}^D \times \left(\mathbb{R}^D\right)^{n-1}$ *are* $\bar{Q}^{(i)}_{\tau,1}$ *and* $\bar{Q}^{(i)}_{\tau,n-1}$, *and we have*

$$\int_{U'_k} d\bar{Q}^{(i)}_{\tau,n-1} \wedge d\bar{Q}^{(i)}_{\tau^k,n-1} \geq c_0, \quad \text{and} \quad m \cdot \int_{U_k} d\bar{Q}^{(i)}_{\tau,1} \wedge d\bar{Q}^{(i)}_{\tau^k,1} \geq c_d.$$

The second family, described in Lemma 9, testifies of the statistical difficulty of the estimation problem when the noise level σ is large enough. The construction is very similar to Lemma 8 (see Figure 6). Though, in this case, the magnitude of the noise drives the statistical difficulty, as opposed to the sampling scale in Lemma 8. Note that in this case, considering mixture distributions is not necessary since the ample-enough noise make bumps that are absolutely continuous with respect to each other. The proof of Lemma 9 can be found in Section C.3.2 of the Appendix.



(a) Flat bump: $\tau_k = 0$.

(b) Linear bump: $\tau_k = 1$, $i = 1$.

(c) Quadratic bump: $\tau_k = 1$, $i = 2$.

Figure 6: Distributions of Lemma 8 in the neighborhood of U_k ($1 \leq k \leq m$). Black curves correspond to the support $M_\tau^{(i)}$ of a distribution of $\mathcal{P}_\tau^{(i)} \subset \mathcal{P}^k$. The area shaded in grey depicts the mixture distribution $\bar{Q}_{\tau,1}^{(i)} \in \overline{Conv}\left(\mathcal{P}_\tau^{(i)}\right)$.

Lemma 9. *Assume that the conditions of Theorem 3 or 5 hold, and that $\sigma \geq C_{k,d,\tau_{min}} \left(1/(n-1)\right)^{k/d}$ for $C_{k,d,\tau_{min}} > 0$ large enough. Given $i \in \{1,2\}$, there exists a collection of 2^m distributions $\left\{\mathbf{P}_\tau^{(i),\sigma}\right\}_{\tau \in \{0,1\}^m} \subset \mathcal{P}^k(\sigma)$ with associated submanifolds $\left\{M_\tau^{(i),\sigma}\right\}_{\tau \in \{0,1\}^m}$, together with pairwise disjoint subsets $\{U_k^\sigma\}_{1 \leq k \leq m}$ of $\mathbb{R}^D$ such that the following holds for all $\tau \in \{0,1\}^m$ and $1 \leq k \leq m$.*

If $x \in U_k^\sigma$ and $y = \pi_{M_\tau^{(i),\sigma}}(x)$, we have

- *if $\tau_k = 0$,*

$$T_y M_\tau^{(i),\sigma} = \mathbb{R}^d \times \{0\}^{D-d} \quad , \quad \left\| II_y^{M_\tau^{(i),\sigma}} \circ \pi_{T_y M_\tau^{(i),\sigma}} \right\|_{op} = 0,$$

- *if $\tau_k = 1$,*
  - *for $i = 1$:* $\angle\left(T_y M_\tau^{(1),\sigma}, \mathbb{R}^d \times \{0\}^{D-d}\right) \geq c_{k,d,\tau_{min}} \left(\dfrac{\sigma}{n-1}\right)^{\frac{k-1}{k+d}}$,

– *for* $i = 2$*:* $\left\| II_y^{M_\tau^{(2),\sigma}} \circ \pi_{T_y M_\tau^{(2),\sigma}} \right\|_{op} \geq c'_{k,d,\tau_{min}} \left(\frac{\sigma}{n-1} \right)^{\frac{k-2}{k+d}}$.

Furthermore,

$$\int_{(\mathbb{R}^D)^{n-1}} \left(\mathbf{P}_\tau^{(i),\sigma}\right)^{\otimes n-1} \wedge \left(\mathbf{P}_{\tau^k}^{(i),\sigma}\right)^{\otimes n-1} \geq c_0, \text{ and } \quad m \cdot \int_{U_k^\sigma} \mathbf{P}_\tau^{(i),\sigma} \wedge \mathbf{P}_{\tau^k}^{(i),\sigma} \geq c_d.$$

Proof of Theorem 3. Let us apply Lemma 7 with $\mathcal{X} = \mathbb{R}^D$, $\mathcal{X}' = \left(\mathbb{R}^D\right)^{n-1}$, $\mathcal{Q} = \left(\mathcal{P}^k(\sigma)\right)^{\otimes n}$, $X = X_1$, $X' = (X_2, \dots, X_n) = X_{2:n}$, $\theta_X(Q) = T_X M$, and the angle between linear subspaces as the distance d.

If $\sigma < C_{k,d,\tau_{min}} \left(1/(n-1)\right)^{k/d}$, for $C_{k,d,\tau_{min}} > 0$ defined in Lemma 9, then, applying Lemma 7 to the family $\left\{\bar{Q}_{\tau,n}^{(1)}\right\}_\tau$ together with the disjoint sets $U_k \times U'_k$ of Lemma 8, we get

$$\begin{aligned} \inf_{\hat{T}} \sup_{P \in \mathcal{P}^k(\sigma)} \mathbb{E}_{P^{\otimes n}} \angle\left(T_{\pi_M(X_1)} M, \hat{T}\right) &\geq m \cdot c_{k,d,\tau_{min}} \left(\frac{1}{n-1}\right)^{\frac{k-1}{d}} \cdot c_0 \cdot c_d \\ &= c'_{d,k,\tau_{min}} \left\{ \left(\frac{1}{n-1}\right)^{\frac{k-1}{d}} \vee \left(\frac{\sigma}{n-1}\right)^{\frac{k-1}{d+k}} \right\}, \end{aligned}$$

where the second line uses that $\sigma < C_{k,d,\tau_{min}} \left(1/(n-1)\right)^{k/d}$.

If $\sigma \geq C_{k,d,\tau_{min}} \left(1/(n-1)\right)^{k/d}$, then Lemma 9 holds, and considering the family $\left\{\left(\mathbf{P}_\tau^{(1),\sigma}\right)^{\otimes n}\right\}_\tau$, together with the disjoint sets $U_k^\sigma \times \left(\mathbb{R}^D\right)^{n-1}$, Lemma 7 gives

$$\begin{aligned} \inf_{\hat{T}} \sup_{P \in \mathcal{P}^k(\sigma)} \mathbb{E}_{P^{\otimes n}} \angle\left(T_{\pi_M(X_1)} M, \hat{T}\right) &\geq m \cdot c_{k,d,\tau_{min}} \left(\frac{\sigma}{n-1}\right)^{\frac{k-1}{k+d}} \cdot c_0 \cdot c_d \\ &= c''_{d,k,\tau_{min}} \left\{ \left(\frac{1}{n-1}\right)^{\frac{k-1}{d}} \vee \left(\frac{\sigma}{n-1}\right)^{\frac{k-1}{d+k}} \right\}, \end{aligned}$$

hence the result.

□

Proof of Theorem 5. The proof follows the exact same lines as that of Theorem 3 just above. Namely, consider the same setting with $\theta_X(Q) = II^M_{\pi_M(X)}$. If $\sigma \geq C_{k,d,\tau_{min}} \left(1/(n-1)\right)^{k/d}$, apply Lemma 7 with the family $\left\{\bar{Q}_{\tau,n}^{(2)}\right\}_\tau$ of Lemma 8. If $\sigma > C_{k,d,\tau_{min}} \left(1/(n-1)\right)^{k/d}$, Lemma 7 can be applied to $\left\{\left(\mathbf{P}_\tau^{(2),\sigma}\right)^{\otimes n}\right\}_\tau$ in Lemma 9. This yields the announced rate. □

Acknowledgements. We would like to thank Frédéric Chazal and Pascal Massart for their constant encouragements, suggestions and stimulating discussions. We also thank the anonymous reviewers for valuable comments and suggestions.

REFERENCES

[1] Aamari, E. and Levrard, C. (2015). Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction. *ArXiv e-prints*.

[2] Aamari, E. and Levrard, C. (2017). Supplementary file for: Non-asymptotic rates for manifold, tangent space and curvature estimation.

[3] Arias-Castro, E., Lerman, G. and Zhang, T. (2013). Spectral Clustering Based on Local PCA. *ArXiv e-prints*.

[4] Arias-Castro, E., Pateiro-López, B. and Rodríguez-Casal, A. (2016). Minimax Estimation of the Volume of a Set with Smooth Boundary. *ArXiv e-prints*.

[5] Boissonnat, J.-D. and Ghosh, A. (2014). Manifold reconstruction using tangential Delaunay complexes. *Discrete Comput. Geom.* **51** 221–267. MR3148657

[6] Cazals, F. and Pouget, M. (2005). Estimating differential quantities using polynomial fitting of osculating jets. *Comput. Aided Geom. Design* **22** 121–146. MR2116098

[7] Cheng, S.-W. and Chiu, M.-K. (2016). Tangent estimation from point samples. *Discrete Comput. Geom.* **56** 505–557. MR3544007

[8] Cheng, S.-W., Dey, T. K. and Ramos, E. A. (2005). Manifold reconstruction from point samples. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 1018–1027. ACM, New York. MR2298361

[9] do Carmo, M. P. (1992). *Riemannian geometry. Mathematics: Theory & Applications.* Birkhäuser Boston, Inc., Boston, MA Translated from the second Portuguese edition by Francis Flaherty. MR1138207 (92i:53001)

[10] Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S. and Singh, A. (2014). Confidence sets for persistence diagrams. *Ann. Statist.* **42** 2301–2339. MR3269981

[11] Federer, H. (1959). Curvature measures. *Trans. Amer. Math. Soc.* **93** 418–491. MR0110078 (22 ##961)

[12] Fefferman, C., Ivanov, S., Kurylev, Y., Lassas, M. and Narayanan, H. (2015). Reconstruction and interpolation of manifolds I: The geometric Whitney problem. *ArXiv e-prints*.

[13] Gashler, M. S. and Martinez, T. (2011). Tangent Space Guided Intelligent Neighbor Finding. In *Proceedings of the IEEE International Joint Conference on Neural Networks IJCNN'11* 2617–2624. IEEE Press.

[14] Genovese, C. R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2012). Manifold estimation and singular deconvolution under Hausdorff loss. *Ann. Statist.* **40** 941–963. MR2985939

[15] Genovese, C. R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2012). Minimax manifold estimation. *J. Mach. Learn. Res.* **13** 1263–1291. MR2930639

[16] Golub, G. H. and Van Loan, C. F. (1996). *Matrix computations*, third ed. *Johns Hopkins Studies in the Mathematical Sciences.* Johns Hopkins University Press, Baltimore, MD. MR1417720 (97g:65006)

[17] Gumhold, S., Wang, X. and MacLeod, R. (2001). Feature Extraction from Point Clouds. In *10th International Meshing Roundtable* 293–305. Sandia National Laboratories,.

[18] Hartman, P. (1951). On geodesic coordinates. *Amer. J. Math.* **73** 949–954. MR0046087

[19] Kim, A. K. H. and Zhou, H. H. (2015). Tight minimax rates for manifold estimation under Hausdorff loss. *Electron. J. Stat.* **9** 1562–1582. MR3376117

[20] Maggioni, M., Minsker, S. and Strawn, N. (2016). Multiscale dictionary learning: non-asymptotic bounds and robustness. *J. Mach. Learn. Res.* **17** Paper No. 2, 51. MR3482922

[21] Merigot, Q., Ovsjanikov, M. and Guibas, L. J. (2011). Voronoi-Based Curvature and Feature Estimation from Point Clouds. *IEEE Transactions on Visualization and Computer Graphics* **17** 743-756.

[22] Niyogi, P., Smale, S. and Weinberger, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.* **39** 419–441. MR2383768 (2009b:60038)

[23] Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE* **290** 2323–2326.

[24] Rusinkiewicz, S. (2004). Estimating Curvatures and Their Derivatives on Triangle Meshes. In *Proceedings of the 3D Data Processing, Visualization, and Transmission, 2Nd International Symposium. 3DPVT '04* 486–493. IEEE Computer Society, Washington, DC, USA.

[25] Singer, A. and Wu, H. T. (2012). Vector diffusion maps and the connection Laplacian. *Comm. Pure Appl. Math.* **65** 1067–1144. MR2928092

[26] Tenenbaum, J. B., de Silva, V. and Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **290** 2319.

[27] Tyagi, H., Vural, E. f. and Frossard, P. (2013). Tangent space estimation for smooth embeddings of Riemannian manifolds. *Inf. Inference* **2** 69–114. MR3311444

[28] Usevich, K. and Markovsky, I. (2014). Optimization on a Grassmann manifold with application to system identification. *Automatica* **50** 1656 - 1662.

[29] Wasserman, L. (2016). Topological Data Analysis. *ArXiv e-prints.*

[30] Yu, B. (1997). Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam* 423–435. Springer.

SUPPLEMENTARY MATERIAL

Appendix: Geometric background and proofs of intermediate results

(doi: COMPLETED BY THE TYPESETTER; .pdf). Due to space constraints, we relegate technical details of the remaining proofs to the supplement [2].

Department of Mathematics
University of California San Diego
9500 Gilman Dr. La Jolla
CA 92093
United States
E-mail: eaamari@ucsd.edu
URL: http://www.math.ucsd.edu/ eaamari/

Laboratoire de probabilités et modèles aléatoires
Bâtiment Sophie Germain
Université Paris-Diderot
75013 Paris
France
E-mail: levrard@math.univ-paris-diderot.fr
URL: http://www.normalesup.org/ levrard/