Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction

Eddie Aamari, Clément Levrard

Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction

Eddie Aamari and Clément Levrard

Abstract

In this paper we consider the problem of optimality in manifold reconstruction. A random sample $\mathbb{X}_n = \{X_1, \dots, X_n\} \subset \mathbb{R}^D$ composed of points lying on a d-dimensional submanifold M, with or without outliers drawn in the ambient space, is observed. Based on the Tangential Delaunay Complex [4], we construct an estimator $\hat{M}$ that is ambient isotopic and Hausdorff-close to M with high probability. The estimator $\hat{M}$ is built from existing algorithms. In a model without outliers, we show that this estimator is asymptotically minimax optimal for the Hausdorff distance over a class of submanifolds defined with a reach constraint. Therefore, even with no *a priori* information on the tangent spaces of M, our estimator based on Tangential Delaunay Complexes is optimal. This shows that the optimal rate of convergence can be achieved through existing algorithms. A similar result is also derived in a model with outliers. A geometric interpolation result is derived, showing that the Tangential Delaunay Complex is stable with respect to noise and perturbations of the tangent spaces. In the process, a denoising procedure and a tangent space estimator both based on local principal component analysis (PCA) are studied.

1 Introduction

Throughout many fields of applied science, data in $\mathbb{R}^D$ can naturally be modeled as lying on a d-dimensional submanifold M. As M may carry a lot of information about the studied phenomenon, it is then natural to consider the problem of either approximating M geometrically, recovering it topologically, or both from a point sample $\mathbb{X}_n = \{X_1, \dots, X_n\}$. It is of particular interest in high codimension ($d \ll D$) where it can be used as a preliminary processing of the data for reducing its dimension, and then avoiding the curse of dimensionality. This problem is usually referred to as *manifold reconstruction* in the computational geometry community, and rather called *set/support estimation* or *manifold learning* in the statistics literature.

The computational geometry community has now been active on manifold reconstruction for many years, mainly in deterministic frameworks. In dimension 3, [16] provides a survey of the state of the art. In higher dimension, the employed methods rely on variants of the ambient Delaunay triangulation [12, 4]. The geometric and topological guarantees are derived under the assumption that the point cloud — fixed and nonrandom — densely samples M at scale ε, with ε small enough or going to 0.

In the statistics literature, most of the attention has been paid to approximation guarantees, rather than topological ones. The approximation bounds are given in terms of the sample size n, that is assumed to be large enough or going to infinity. To derive these bounds, a broad variety of assumptions on M have been considered. For instance, if M is a bounded convex set and $\mathbb{X}_n$ does not contain outliers, a natural idea is to consider the convex hull $\hat{M} = Conv(\mathbb{X}_n)$ to be the

estimator. $Conv(\mathbb{X}_n)$ provides optimal rates of approximation for several loss functions [28, 19]. These rates depend crudely on the regularity of the boundary of the convex set M. In addition, $Conv(\mathbb{X}_n)$ is clearly ambient isotopic to M so that it has both good geometric and topological properties. Generalisations of the notion of convexity based on rolling ball-type assumptions such as r-convexity and reach bounds [14, 23] yield rich classes of sets with good geometric properties. In particular, the reach, as introduced by Federer [21], appears to be a key regularity and scale parameter [11, 23].

This paper mainly follows up the two articles [4, 23], both dealing with the case of a d-dimensional manifold $M \subset \mathbb{R}^D$ with a reach regularity condition and where the dimension d is known.
On one hand, [4] focuses on a deterministic analysis and proposes a provably faithful reconstruction. The authors introduce a weighted Delaunay triangulation restricted to tangent spaces, the so-called Tangential Delaunay Complex. This paper gives a reconstruction up to ambient isotopy with approximation bounds for the Hausdorff distance along with computational complexity bounds. This work provides a simplicial complex based on the input point cloud and tangent spaces. However, it lacks stability up to now, in the sense that the assumptions used in the proofs of [4] do not resist ambient perturbations. Indeed, it heavily relies on the knowledge of the tangent spaces at each point and on the absence of noise.
On the other hand, [23] takes a statistical approach in a model possibly containing outlier points. The authors derive an estimator that is proved to be minimax optimal for the Hausdorff distance $\mathrm{d_H}$. Roughly speaking, minimax optimality of the proposed estimator means that it performs best in the worst possible case up to numerical constants, when the sample size n is large enough. Although theoretically optimal, the proposed estimator appears to be intractable in practice.

Our main contribution (Theorem 6) makes a two-way link between the approaches of [4] and [23]. It shows that with an additional tangent space estimation procedure that we propose, the Tangential Delaunay Complex of [4] achieves the optimal rate of convergence in the model without outliers of [23]. Conversely, Theorem 6 points out that the rate of [23] can be achieved with a tractable estimator $\hat{M}$ that is a simplicial complex of which vertices are the data points and such that $\hat{M}$ is ambient isotopic to M with high probability. In the presence of outliers, an iterative denoising procedure is proposed, and similar reconstruction results (Theorem 7 and Theorem 8) are derived for a simplicial complex based on the denoised point cloud. Moreover, a stability result for the Tangential Delaunay Complex (Theorem 13) is proved.

Outline

This paper deals with the case where a sample $\mathbb{X}_n = \{X_1, \dots, X_n\} \subset \mathbb{R}^D$ of size n is randomly drawn on/around M. First, the statistical framework is described (Section 2). Two models are studied, one where $\mathbb{X}_n \subset M$ and the other where $\mathbb{X}_n$ contains outliers. We build a simplicial complex $\hat{M}_{\texttt{TDC}}(\mathbb{X}_n)$ ambient isotopic to M and achieving a rate of approximation for the Hausdorff distance $\mathrm{d_H}(M, \hat{M}_{\texttt{TDC}})$, with bounds holding uniformly over a class of submanifolds satisfying a reach regularity condition. The derived rate of convergence is minimax optimal (Section 2) in the model without outliers. With outliers,similar estimators $\hat{M}_{\texttt{TDC}\delta}$ and $\hat{M}_{\texttt{TDC}+}$ are built. $\hat{M}_{\texttt{TDC}}$, $\hat{M}_{\texttt{TDC}\delta}$ and $\hat{M}_{\texttt{TDC}+}$ are based on the Tangential Delaunay Complex (Section 3), that is first proved to be stable (Section 4) via an interpolation result. For this purpose, a method to estimate tangent spaces and to remove outliers based on local Principal Component Analysis (PCA) is proposed (Section 5). We conclude with general remarks and possible extensions (Section 6). For ease of exposition, all

the proofs are placed in the appendix.

Notation

In what follows, we consider a smooth d-dimensional submanifold without boundary $M \subset \mathbb{R}^D$ to be reconstructed. For all $p \in M$, T_pM designates the tangent space of M at p. Tangent spaces will either be considered vectorial or affine depending on the context. The standard inner product in $\mathbb{R}^D$ is denoted by $\langle \cdot, \cdot \rangle$ and the Euclidean distance $\|\cdot\|$. We let $\mathcal{B}(p, r)$ denote the closed Euclidean ball of radius $r > 0$ centered at p. As introduced in [21], the reach of M, denoted by $\mathrm{reach}(M)$ is the maximal offset radius for which the projection π_M onto M is well defined. Denoting by $\mathrm{d}(\cdot, M)$ the distance to M, the *medial axis* of M $\mathrm{med}(M) = \{x \in \mathbb{R}^D | \exists a \neq b \in M, \|x-a\| = \|x-b\| = \mathrm{d}(x, M)\}$ is the set of points which have at least two nearest neighbors on M. Then, $\mathrm{reach}(M) = \inf_{p \in M} \mathrm{d}(p, \mathrm{med}(M))$. We simply write π for π_M when there is no possibility of confusion. For any smooth function $\Phi : \mathbb{R}^D \to \mathbb{R}^D$, we let $\mathrm{d}_a\Phi$ and $\mathrm{d}_a^2\Phi$ denote the first and second order differentials of Φ at $a \in \mathbb{R}^D$. For a linear map A, A^t designates its transpose. Let $\|A\|_{\mathrm{op}} = \sup_x \frac{\|Ax\|}{\|x\|}$ and $\|A\|_{\mathcal{F}} = \sqrt{\mathrm{trace}\,(A^t A)}$ denote respectively the operator norm induced by the Euclidean norm and the Frobenius norm. The distance between two linear subspaces $U, V \subset \mathbb{R}^D$ of the same dimension is measured by the principal angle $\angle(U, V) = \max_{u \in U} \max_{v' \in V^\perp} \frac{\langle u, v' \rangle}{\|u\| \, \|v'\|} = \|\pi_U - \pi_V\|_{\mathrm{op}}$. The Hausdorff distance in $\mathbb{R}^D$ is denoted by $\mathrm{d_H}$. We let $\cong$ denote the ambient isotopy relation in $\mathbb{R}^D$. Throughout this paper, C_α will denote a generic constant depending on the parameter α. For clarity's sake, c_α and K_α may also be used when several constants are involved.

2 Minimax Risk and Optimality

2.1 Statistical Model

Let us describe the general statistical setting we will use to define optimality for manifold reconstruction. A *statistical model* $\mathcal{D}$ is a set of probability distributions on $\mathbb{R}^D$. In any statistical experiment, $\mathcal{D}$ is fixed and known. We observe an independent and identically distributed sample of size n (or i.i.d. n-sample) $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ drawn according to some unknown distribution $P \in \mathcal{D}$. If no noise is allowed, the problem is to recover the *support* of P, that is, the smallest closed set $C \subset \mathbb{R}^D$ such that $P(C) = 1$. Let us give two examples of such models $\mathcal{D}$ by describing those of interest in this paper.

Let $\mathcal{M}_{D,d,\rho}$ be the set of all the d-dimensional connected submanifolds $M \subset \mathbb{R}^D$ without boundary satisfying $\mathrm{reach}(M) \geq \rho$. The reach assumption is crucial to avoid arbitrarily curved and pinched shapes [14]. From a reconstruction point of view, ρ gives a minimal feature size on M, and then a minimal scale for geometric information. Every $M \in \mathcal{M}_{D,d,\rho}$ inherits a measure induced by the d-dimensional Hausdorff measure on $\mathbb{R}^D \supset M$. We denote this induced measure v_M. Beyond the geometric restrictions induced by the lower bound ρ on the reach, it also requires the natural measure v_M to behave like a d-dimensional measure, up to uniform constants. Namely, v_M satisfies the (a, d)-standard property of [10], with $a = a_{d,\rho}$. Denote by $\mathcal{U}_M(f_{min}, f_{max})$ the set of probability distributions Q having a density f with respect to v_M such that $0 < f_{min} \leq f(x) \leq f_{max} < \infty$ for all $x \in M$. Roughly speaking, when $Q \in \mathcal{U}_M(f_{min}, f_{max})$, points are drawn almost uniformly

on M . This is to ensure that the sample visits all the areas of M with high probability. The model without outliers $\mathcal{G}_{D,d,f_{min},f_{max},\rho}$ consists of the set of all these almost uniform measures on submanifolds of dimension d having reach greater than a fixed value $\rho > 0$.

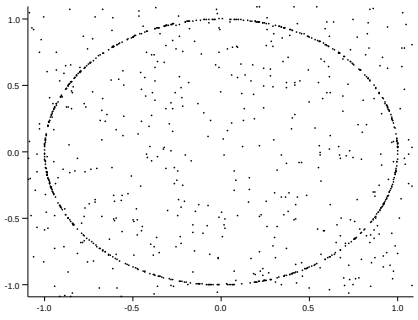
Definition 1 (Model without outliers)**.** $\mathcal{G}_{D,d,f_{min},f_{max},\rho} = \bigcup_{M \in \mathcal{M}_{D,d,\rho}} \mathcal{U}_M(f_{min}, f_{max})$.

We do not explicitly impose a bound on the diameter of M. Actually, a bound is implicitly present in the model, as stated in the next lemma, the proof of which follows from a volume argument.

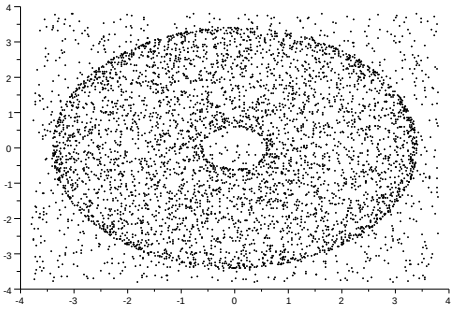
Lemma 2. *There exists $C_d > 0$ such that for all $Q \in \mathcal{G}_{D,d,f_{min},f_{max},\rho}$ with associated M,*

$$\operatorname{diam}(M) \leq \frac{C_d}{\rho^{d-1} f_{min}} =: K_{d,f_{min},\rho}.$$

Eventually, we may include distributions contaminated with outliers uniformly drawn in a ball $\mathcal{B}_0$ containing M, as illustrated in Figure 1. Up to translation, we can always assume that $M \ni 0$. To avoid boundary effects, $\mathcal{B}_0$ will be taken to contain M amply, so that the outlier distribution surrounds M everywhere. Since M has at most diameter $K_{d,f_{min},\rho}$ from Lemma 2 we arbitrarily fix $\mathcal{B}_0 = \mathcal{B}(0, K_0)$, where $K_0 = K_{d,f_{min},\rho} + \rho$. Notice that the larger the radius of $\mathcal{B}_0$, the easier to label the outlier points since they should be very far away from each other.



(a) Circle with outliers: $d = 1$,$D = 2$.



(b) Torus with outliers: $d = 2$,$D = 3$.

Figure 1: Point clouds $\mathbb{X}_n$ drawn from distributions of two instances of models $\mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}$ with different parameters and $\beta < 1$.

Definition 3 (Model with outliers/Clutter noise model)**.** *For $0 < f_{min} \leq f_{max} < \infty$, $0 < \beta \leq 1$, and $\rho > 0$, we define $\mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}$ to be the set of mixture distributions*

$$P = \beta Q + (1 - \beta) U_{\mathcal{B}_0},$$

where $Q \in \mathcal{G}_{D,d,f_{min},f_{max},\rho}$ has support M such that $0 \in M$, and $U_{\mathcal{B}_0}$ is the uniform distribution on $\mathcal{B}_0 = \mathcal{B}(0, K_0)$.

Alternatively, a random variable X with distribution $P \in \mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}$ can be represented as $X = VX' + (1 - V)X''$, where $V \in \{0, 1\}$ is a Bernoulli random variable with parameter β, X' has distribution in $\mathcal{G}_{D,d,f_{min},f_{max},\rho}$ and X'' has a uniform distribution over $\mathcal{B}_0$, and such that V, X', X'' are independent. In particular for $\beta = 1$, $\mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta=1} = \mathcal{G}_{D,d,f_{min},f_{max},\rho}$.

2.2 Minimax Risk

For a probability measure $P \in \mathcal{D}$, denote by $\mathbb{E}_P$ — or simply $\mathbb{E}$ — the expectation with respect to the product measure $P^{(n)}$. The quantity we will be interested in is the *minimax risk* associated to the model $\mathcal{D}$. For $n \geq 0$,

$$R_n(\mathcal{D}) = \inf_{\hat{M}} \sup_{P \in \mathcal{D}} \mathbb{E}_P \left[\mathrm{d_H} \left(M, \hat{M} \right) \right],$$

where the infimum is taken over all the estimators $\hat{M} = \hat{M}(X_1, \ldots, X_n)$ computed over an n-sample. $R_n(\mathcal{D})$ is the best risk that an estimator based on an n-sample can achieve uniformly over the class $\mathcal{D}$. It is clear from the definition that if $\mathcal{D}' \subset \mathcal{D}$ then $R_n(\mathcal{D}') \leq R_n(\mathcal{D})$. It follows the intuition that the broader the class of considered manifolds, the more difficult it is to estimate them uniformly well. Studying $R_n(\mathcal{D})$ for a fixed n is a difficult task that can rarely be carried out. We will focus on the semi-asymptotic behavior of this risk. As $R_n(\mathcal{D})$ cannot be surpassed, its rate of convergence to 0 as $n \to \infty$ may be seen as the best rate of approximation that an estimator can achieve. More precisely, we will say that two sequences $(a_n)_n$ and $(b_n)_n$ are asymptotically comparable, denoted by $a_n \asymp b_n$, if there exist $c, C > 0$ such that for n large enough, $cb_n \leq a_n \leq Cb_n$.

Definition 4. *An estimator $\hat{M}$ is said to be (asymptotically) minimax optimal over $\mathcal{D}$ if*

$$\sup_{P \in \mathcal{D}} \mathbb{E}_P \left[\mathrm{d_H} \left(M, \hat{M} \right) \right] \asymp R_n(\mathcal{D}).$$

In other words, $\hat{M}$ is (asymptotically) minimax optimal if it achieves, up to constants, the best possible rate of convergence in the worst case.

Studying a minimax rate of convergence is twofold. On one hand, deriving an upper bound on R_n boils down to give an estimator and to study its quality uniformly on $\mathcal{D}$. On the other hand, lower bounding R_n amounts to study the worst possible case in $\mathcal{D}$. This part is usually achieved with standard Bayesian techniques [26]. For the models considered in the present paper, the rates were given in [23, 25].

Theorem 5 (Theorem 3 of [25])**.** $R_n\left(\mathcal{G}_{D,d,f_{min},f_{max},\rho}\right) \asymp R_n\left(\mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}\right) \asymp \left(\frac{\log n}{\beta n}\right)^{2/d}$.

Beyond this theoretical result, an interesting question is to know whether this minimax rate can be achieved by a tractable algorithm. Indeed, that proposed in [23] especially relies on a minimization problem over the class of submanifolds $\mathcal{M}_{D,d,\rho}$, which seems computationally costly. In addition, the proposed estimator is itself a manifold, that raises storage problems. Moreover, no guarantee is given on the topology of the estimator.

Throughout the present paper, we will build an estimator $\hat{M}_{\mathtt{TDC}}$ that addresses these issues. $\hat{M}_{\mathtt{TDC}}$ is based on the Tangential Delaunay Complex (Section 3), with a tangent space estimation using a local PCA (Section 5). The result, derived in the model without outliers $\mathcal{G}_{D,d,f_{min},f_{max},\rho}$, is stated as follows.

Theorem 6. *$\hat{M}_{\mathtt{TDC}} = \hat{M}_{\mathtt{TDC}}(\mathbb{X}_n)$ is a simplicial complex of vertices $\mathbb{X}_n$ such that*

$$\lim_{n \to \infty} \mathbb{P}\left(\mathrm{d_H}(M, \hat{M}_{\mathtt{TDC}}) \leq C_{d,f_{min},f_{max},\rho} \left(\frac{\log(n-1)}{n-1} \right)^{2/d} \text{ and } M \cong \hat{M}_{\mathtt{TDC}} \right) = 1.$$

Moreover, $\hat{M}_{\mathtt{TDC}}$ is minimax optimal over $\mathcal{G}_{D,d,f_{min},f_{max},\rho}$: for n large enough,

$$\sup_{Q\in\mathcal{G}_{D,d,f_{min},f_{max},\rho}} \mathbb{E}_Q \mathrm{d_H}(M,\hat{M}_{\texttt{TDC}}) \leq C'_{d,f_{min},f_{max},\rho}\left(\frac{\log(n-1)}{n-1}\right)^{2/d}.$$

It is interesting to note that the constants appearing in Theorem 6 do not depend on the ambient dimension D.

Furthermore, in the model with outliers $\mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}$, with the same procedure used to derive Theorem 6 and an additional iterative preprocessing of the data based on local PCA to remove outliers (Section 5), we can design an estimator of M that is as close as wanted to the noise-free case estimator. Namely, for any positive $\delta < 1/(d(d+1))$, we build $\hat{M}_{\texttt{TDC}\delta}$ that satisfies the following similar statement.

Theorem 7. *$\hat{M}_{\texttt{TDC}\delta} = \hat{M}_{\texttt{TDC}\delta}(\mathbb{X}_n)$ is a simplicial complex of vertices contained in $\mathbb{X}_n$ such that*

$$\lim_{n\to\infty}\mathbb{P}\left(\mathrm{d_H}(M,\hat{M}_{\texttt{TDC}\delta}) \leq C_{d,f_{min},f_{max},\rho}\left(\frac{\log(n-1)}{\beta(n-1)}\right)^{2/d-2\delta} \text{ and } M\cong\hat{M}_{\texttt{TDC}\delta}\right)=1.$$

Moreover, for n large enough,

$$\sup_{P\in\mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}} \mathbb{E}_Q \mathrm{d_H}(M,\hat{M}_{\texttt{TDC}+}) \leq C'_{d,f_{min},f_{max},\rho}\left(\frac{\log(n-1)}{\beta(n-1)}\right)^{2/d-2\delta}.$$

$\hat{M}_{\texttt{TDC}\delta}$ converges at the rate at least $\left(\frac{\log n}{n}\right)^{2/d-2\delta}$, which is not the minimax optimal rate according to Theorem 5 but can be set as close as desired to it. To our knowledge, $\hat{M}_{\texttt{TDC}\delta}$ is the first explicit estimator to provably achieve such a rate in the presence of outliers. Again, it is worth noting that the constants involved in Theorem 7 do not depend on the ambient dimension D.

In fact, $\hat{M}_{\texttt{TDC}\delta}$ is based on a finite number of iterations of a denoising process. From a theoretical point of view, there exists a (random) number of iterations from which an estimator $\hat{M}_{\texttt{TDC}+}$ can be built to satisfy the following.

Theorem 8. *$\hat{M}_{\texttt{TDC}+} = \hat{M}_{\texttt{TDC}+}(\mathbb{X}_n)$ is a simplicial complex of vertices contained in $\mathbb{X}_n$ such that*

$$\lim_{n\to\infty}\mathbb{P}\left(\mathrm{d_H}(M,\hat{M}_{\texttt{TDC}+}) \leq C_{d,f_{min},f_{max},\rho}\left(\frac{\log(n-1)}{\beta(n-1)}\right)^{2/d} \text{ and } M\cong\hat{M}_{\texttt{TDC}+}\right)=1.$$

Moreover, for n large enough,

$$\sup_{P\in\mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}} \mathbb{E}_Q \mathrm{d_H}(M,\hat{M}_{\texttt{TDC}+}) \leq C'_{d,f_{min},f_{max},\rho}\left(\frac{\log(n-1)}{\beta(n-1)}\right)^{2/d}.$$

$\hat{M}_{\texttt{TDC}+}$ may be thought of as a limit of $\hat{M}_{\texttt{TDC}\delta}$ when δ goes to 0. As it will be proved in Section 5, this limit will be reached for δ close enough to 0. Unfortunately this convergence threshold is also random, hence unknown.

The statistical analysis of the reconstruction problem is postponed to Section 5. Beforehand, let us describe the Tangential Delaunay Complex in a deterministic and idealized framework where the tangent spaces are known and no outliers are present.

3 Tangential Delaunay Complex

Let $\mathcal{P} = \{p_1, \ldots, p_n\} \subset M$. In this section, we denote the point cloud $\mathcal{P}$ to emphasize the fact that it is considered nonrandom. For $\varepsilon, \delta > 0$, $\mathcal{P}$ is said to be *ε-dense* in M if $\mathrm{d_H}(M, \mathcal{P}) \leq \varepsilon$, and *$\delta$-sparse* if $\mathrm{d}(p, \mathcal{P} \setminus \{p\}) \geq \delta$ for all $p \in \mathcal{P}$. A *(δ, ε)-net* (of M) is a δ-sparse and ε-dense point cloud.

3.1 Restricted Weighted Delaunay Triangulations

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

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

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

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

Note that for a constant weight assignment $\omega(p) = \omega_0$, $\mathrm{Del}^\omega(\mathcal{P})$ is the usual Delaunay triangulation of $\mathcal{P}$. Under genericity assumptions on $\mathcal{P}$ and bounds on ω, $\mathrm{Del}^\omega(\mathcal{P})$ is an embedded triangulation with vertex set $\mathcal{P}$ [4]. The reconstruction method proposed in this paper is based on $\mathrm{Del}^\omega(\mathcal{P})$ for some weights ω to be chosen later. As it is a triangulation of the whole convex hull of $\mathcal{P}$ and fails to recover the geometric structure of M, we take restrictions of it in the following manner.

Given a family $R = \{R_p\}_{p \in \mathcal{P}}$ of subsets $R_p \subset \mathbb{R}^D$ indexed by $\mathcal{P}$, the weighted Delaunay complex restricted to R is the sub-complex of $\mathrm{Del}^\omega(\mathcal{P})$ defined by

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

In particular, we define the *Tangential Delaunay Complex* $\mathrm{Del}^\omega(\mathcal{P}, T)$ by taking $R = T = \{T_p M\}_{p \in \mathcal{P}}$, the family of tangent spaces taken at the points of $\mathcal{P} \subset M$ [4]. $\mathrm{Del}^\omega(\mathcal{P}, T)$ is a pruned version of $\mathrm{Del}^\omega(\mathcal{P})$ where only the simplices with directions close to the tangent spaces are kept. Indeed, $T_p M$ being the best linear approximation of M at p, it is very unlikely for a reconstruction of M to have components in directions normal to $T_p M$ — see Figure 2. As pointed out in [4], computing $\mathrm{Del}^\omega(\mathcal{P}, T)$ only requires to compute Delaunay triangulations in the tangent spaces that have dimension d. This reduces the computational complexity dependency on the ambient dimension $D > d$. The weight assignment ω gives degrees of freedom for the reconstruction. The extra degree of freedom ω permits to stabilize the triangulation and to remove the so-called *inconsistencies*, the points remaining fixed. For further details, see [5, 4].

3.2 Guarantees

The following result sums up the reconstruction properties of the Tangential Delaunay Complex that we will use. For more details about it, the reader is referred to [4].

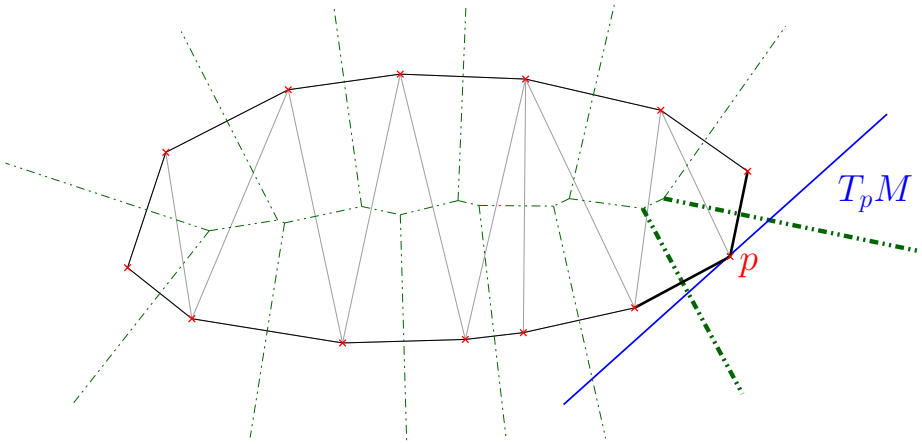


Figure 2: Construction of $\mathrm{Del}^{\omega}(\mathcal{P}, T)$ at p for $\omega \equiv 0$.

Theorem 9 (Theorem 5.3 in [4]). *There exists ε_0 such that for all $\varepsilon \leq \varepsilon_0$ and all $M \in \mathcal{M}_{D,d,\rho}$, if $\mathcal{P} \subset M$ is an $(\varepsilon, 2\varepsilon)$-net, there exists a weight assignment $\omega_* = \omega_{*\mathcal{P},T,}$ depending on $\mathcal{P}$ and $T = \{T_pM\}_{p\in\mathcal{P}}$ such that*

- $\mathrm{d_H}\left(M, \mathrm{Del}^{\omega_*}(\mathcal{P}, T)\right) \leq C_{d,\rho}\varepsilon^2$,
- *M and $\mathrm{Del}^{\omega_*}(\mathcal{P}, T)$ are ambient isotopic.*

Computing $\mathrm{Del}^{\omega_*}(\mathcal{P}, T)$ requires to determine the weight function $\omega_* = \omega_{*\mathcal{P},T}$. In [4], a greedy algorithm is designed for this purpose and has a time complexity $O\left(Dn^2 + D2^{O(d^2)}n\right)$.

Given an $(\varepsilon, 2\varepsilon)$-net $\mathcal{P}$ for ε small enough, $\mathrm{Del}^{\omega_*}(\mathcal{P}, T)$ recovers M up to ambient isotopy and approximates it at the scale ε^2. The order of magnitude ε^2 with an input $\mathcal{P}$ of scale ε is remarkable. Another instance of this phenomenon is present in [13] in codimension 1. We will show that this ε^2 provides the minimax rate of approximation when dealing with random samples. Therefore, it can be thought of as optimal. Theorem 9 suffers two major imperfections. First, it requires the knowledge of the tangent spaces at each sample point — since $\omega_* = \omega_{*\mathcal{P},T}$ — and no guarantee remains if only approximate tangent spaces are known. Second, the points are assumed to lie exactly on the manifold M, and no noise is allowed. The analysis of $\mathrm{Del}^{\omega_*}(\mathcal{P}, T)$ is sophisticated [4]. Rather than redo the whole study with milder assumptions, we tackle this question with an approximation theory approach (Theorem 10). Instead of studying if $\mathrm{Del}^{\omega_*}(\mathcal{P}', T')$ is stable when $\mathcal{P}'$ lies close to M and T' close to T, we examine what $\mathrm{Del}^{\omega_*}(\mathcal{P}', T')$ actually reconstructs, as detailed in Section 4.

3.3 On the Sparsity Assumption

In Theorem 9, $\mathcal{P}$ is assumed to be dense enough so that it covers all the areas of M. It is also supposed to be sparse at the same scale as the density parameter ε. Indeed, arbitrarily accumulated points would generate non-uniformity and instability for $\mathrm{Del}^{\omega_*}(\mathcal{P}, T)$ [5, 4]. At this stage, we emphasize that the construction of a $(\varepsilon, 2\varepsilon)$-net can be carried out with an ε-dense sample with the following procedure. Given an ε-dense sample $\mathcal{P}$, the *farthest point sampling* algorithm prunes

$\mathcal{P}$ and outputs an $(\varepsilon, 2\varepsilon)$-net $\mathcal{Q} \subset \mathcal{P}$ of M as follows. Initialize at $\mathcal{Q} = \{p_1\} \subset \mathcal{P}$, and while $\max_{p\in\mathcal{P}} \mathrm{d}(p, \mathcal{Q}) > \varepsilon$, add to $\mathcal{Q}$ the farthest point to $\mathcal{Q}$ in $\mathcal{P}$, i.e., $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{\operatorname{argmax}_{p\in\mathcal{P}} \mathrm{d}(p, \mathcal{Q})\}$. The output $\mathcal{Q}$ is ε-sparse and satisfies $\mathrm{d_H}(\mathcal{P}, \mathcal{Q}) \leq \varepsilon$, so it is a $(\varepsilon, 2\varepsilon)$-net of M. Therefore, up to the multiplicative constant 2, sparsifying $\mathcal{P}$ at scale ε will not deteriorate its density property. Then, we can run the farthest point sampling algorithm to preprocess the data, so that the obtained point cloud is a net.

4 Stability Result

4.1 Interpolation Theorem

As mentioned above, if the data do not lie exactly on M and if we do not have the exact knowledge of the tangent spaces, Theorem 9 does not apply. To bypass this issue, we interpolate the data with another manifold M' satisfying good properties, as stated in the following result.

Theorem 10 (Interpolation)**.** *Let $M \in \mathcal{M}_{D,d,\rho}$. Let $\mathcal{P} = \{p_1, \ldots, p_q\} \subset \mathbb{R}^D$ be a finite point cloud and $\tilde{T} = \left\{\tilde{T}_1, \ldots, \tilde{T}_q\right\}$ be a family of d-dimensional linear subspaces of $\mathbb{R}^D$. For $\theta \leq \pi/64$ and $18\eta < \delta \leq \rho$, assume that*

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

Then, there exist universal constants $c_1 \leq 11, c_2 \leq 252$ and a smooth submanifold $M' \subset \mathbb{R}^D$ such that

1. *$\mathcal{P} \subset M'$,*
2. *$\operatorname{reach}(M') \geq \dfrac{\rho \cdot \left(1 - c_1\left(\frac{\eta}{\delta} + \theta\right)\right)^2}{1 + c_1\left(\frac{\eta}{\delta} + \theta\right) + c_2\left(\frac{\eta}{\delta^2} + \frac{\theta}{\delta}\right)\rho}$,*
3. *$T_{p_j}M' = \tilde{T}_j$ for all $1 \leq j \leq q$,*
4. *$\mathrm{d_H}(M, M') \leq \delta\theta + \eta$,*
5. *M and M' are ambient isotopic.*

Theorem 10 fits a manifold M' to noisy points and perturbed tangent spaces with no change of topology and a controlled reach loss. We will use M' as a proxy for M. Indeed, if $\tilde{T}_1, \ldots, \tilde{T}_q$ are estimated tangent spaces at the noisy base points $p_1, \ldots, p_q$, M' has the virtue of being reconstructed by $\mathrm{Del}^{\omega_*}(\mathcal{P}, \tilde{T})$ from Theorem 9. Since M' is topologically and geometrically close to M, we conclude that M is reconstructed as well by transitivity. In other words, Theorem 10 allows one to consider a noisy sample with estimated tangent spaces as an exact sample with exact tangent spaces. M' is built pushing and rotating M towards the p_j's locally along the vector $(p_j - \pi(p_j))$, as illustrated in Figure 3. Since the construction is quite general and may be applied in various settings, let us provide an outline of the construction.

Let $\phi(x) = \exp\left(\frac{\|x\|^2}{\|x\|^2-1}\right)\mathbb{1}_{\|x\|^2<1}$. ϕ is smooth and satisfies $\phi(0) = 1$, $\|\phi\|_\infty \leq 1$ and $\mathrm{d}_0\phi = 0$. For $j = 1, \ldots, q$, it follows easily from the definition of $\angle(T_{\pi(p_j)}M, \tilde{T}_j)$ — *e.g.* by induction on

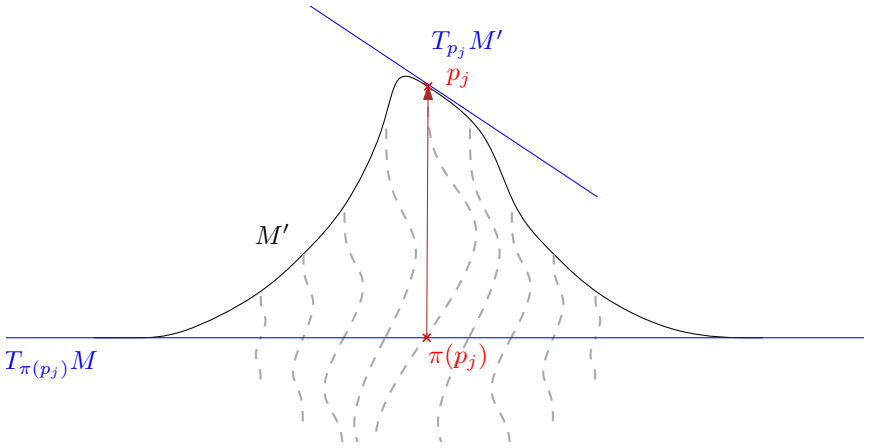


Figure 3: An instance of the interpolating manifold M'. Dashed lines correspond to the image of vertical lines by the ambient diffeomorphism Φ defining $M' = \Phi(M)$.

the dimension — that there exists a rotation R_j of $\mathbb{R}^D$ mapping $T_{\pi(p_j)}M$ onto $\tilde{T}_j$ that satisfies $\|R_j - I_D\|_{op} \leq \theta$. For $\ell > 0$ to be chosen later, and all $a \in \mathbb{R}^D$, let us define $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ by

$$\Phi(a) = a + \sum_{j=1}^{q} \phi\left(\frac{a - \pi(p_j)}{\ell}\right) \underbrace{\left[(R_j - I_D)(a - \pi(p_j)) + (p_j - \pi(p_j))\right]}_{\psi_j(a)}.$$

Φ is designed to map $\pi(p_j)$ onto p_j with $\mathrm{d}_{\pi(p_j)}\Phi = R_j$. Roughly speaking, in balls of radii ℓ around each $\pi(p_j)$, Φ shifts the points in the direction $p_j - \pi(p_j)$ and rotates it around $\pi(p_j)$. Off these balls, Φ is the identity map. To guarantee smoothness, the shifting and the rotation are modulated by the kernel ϕ, as $\|a - \pi(p_j)\|$ increases. Notice that $\mathrm{d}_a\psi_j = (R_j - I_D)$ and $\|\psi_j(a)\| \leq \ell\theta + \eta$ whenever $\phi\left(\frac{a-\pi(p_j)}{\ell}\right) \neq 0$. Defining $M' = \Phi(M)$, the facts that M' fits to $\mathcal{P}$ and $\tilde{T}$ and is Hausdorff-close to M follow by construction. Moreover, Theorem 4.19 of [21] (reproduced as Lemma 23 in this paper) states that the reach is stable with respect to $\mathcal{C}^2$-diffeomorphisms of the ambient space. The estimate on reach(M') relies on the following lemma stating differentials estimates on Φ.

Lemma 11. *There exist universal constants $C_1 \leq 7/2$ and $C_2 \leq 28$ such that if $6\eta < \ell \leq \delta/3$ and $\theta \leq \pi/64$, $\Phi : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ is a global $\mathcal{C}^\infty$-diffeomorphism. In addition, Φ satisfies for all a in $\mathbb{R}^D$,*

$$\|\mathrm{d}_a\Phi\|_{\mathrm{op}} \leq 1 + C_1\left(\frac{\eta}{\ell} + \theta\right), \quad \left\|\mathrm{d}_a\Phi^{-1}\right\|_{\mathrm{op}} \leq \frac{1}{1 - C_1\left(\frac{\eta}{\ell} + \theta\right)}, \quad \left\|\mathrm{d}_a^2\Phi\right\|_{\mathrm{op}} \leq C_2\left(\frac{\eta}{\ell^2} + \frac{\theta}{\ell}\right).$$

The ambient isotopy follows easily by considering the weighted version $\Phi_{(t)}(a) = a + t\,(\Phi(a) - a)$ for $0 \leq t \leq 1$ and the same differential estimates. We then take the maximum possible value $\ell = \delta/3$ and $M' = \Phi(M)$.

Remark 12. *Changing slightly the construction of M', one can also build it such that the curvature tensor at each p_j corresponds to that of M at $\pi(p_j)$. For this purpose it suffices to take a localizing function ϕ identically equal to 1 in a neighborhood of 0. This additional condition would impact the universal constants appearing in Theorem 10.*

4.2 Stability of the Tangential Delaunay Complex

Theorem 10 shows that even in the presence of outliers at distance η from M, and with the knowledge of the tangent spaces up to some angle θ, it is still possible to apply Theorem 9 to some virtual manifold M'. Denoting $\tilde{M} = \mathrm{Del}^{\omega_*}(\mathcal{P}, \tilde{T})$, since $\mathrm{d_H}(M, \tilde{M}) \leq \mathrm{d_H}(M, M') + \mathrm{d_H}(M', \tilde{M})$ and since the ambient isotopy relation is transitive, $M \cong M' \cong \tilde{M}$. We get the following result as a straightforward combination of Theorem 9 and Theorem 10.

Theorem 13 (Stability of the Tangential Delaunay Complex)**.** *There exists $\varepsilon_{d,\rho,\tau_1,\tau_2} > 0$ such that for all $\varepsilon \leq \varepsilon_{d,\rho,\tau_1,\tau_2}$ and all $M \in \mathcal{M}_{D,d,\rho}$, the following holds. Given any finite point cloud $\mathcal{P} \subset \mathbb{R}^D$ and a family $\tilde{T} = \left\{\tilde{T}_p\right\}_{p\in\mathcal{P}}$ of d-dimensional linear subspaces of $\mathbb{R}^D$ such that*

- $\max_{p\in\mathcal{P}} \mathrm{d}(p, M) \leq \tau_1\varepsilon^2$,
- $\mathcal{P}$ *is ε-sparse,*
- $\max_{p\in\mathcal{P}} \angle(T_{\pi(p)}M, \tilde{T}_p) \leq \tau_2\varepsilon$,
- $\max_{x\in M} \mathrm{d}(x, \mathcal{P}) \leq 2\varepsilon$,

then,

- $\mathrm{d_H}\left(M, \mathrm{Del}^{\omega_*}(\mathcal{P}, \tilde{T})\right) \leq \left(C_{d,\rho'_{\rho,\tau_1,\tau_2}} + \tau_1 + \tau_2\right)\varepsilon^2$,
- M *and* $\mathrm{Del}^{\omega_*}(\mathcal{P}, \tilde{T})$ *are ambient isotopic.*

Indeed, applying the reconstruction algorithm of Theorem 9 even in the presence of noise and uncertainty on the tangent spaces actually recovers the manifold M' built in Theorem 10. M' is isotopic to M and the quality of the approximation of M is at most impacted by the term $\mathrm{d_H}(M, M') \leq (\tau_1 + \tau_2)\varepsilon^2$. The lower bound on $\mathrm{reach}(M')$ is crucial, as constants appearing in Theorem 9 are not bounded for arbitrarily small reach ρ'. It is worth noting that no extra analysis of the Tangential Delaunay Complex was needed to derive its stability. The argument is global, constructive, and may be applied to other reconstruction methods taking tangent spaces as input.

5 Tangent Space Estimation and Denoising Procedure

5.1 Noise-Free Case

We now focus on the estimation of tangent spaces in the model without outliers $\mathcal{G}_{D,d,f_{min},f_{max},\rho}$. The proposed method is similar to that of [2]. A point $p \in M$ being fixed, T_pM is the best local d-dimensional linear approximation of M at p. Performing a Local Principal Component Analysis (PCA) in a neighborhood of p would recover the main directions spanned by M at p, and therefore yield a good approximation of T_pM. For $j = 1, \ldots, n$ and $h > 0$ to be chosen later, define the local covariance matrix at X_j by

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

where $N_j = |\mathcal{B}(X_j, h) \cap \mathbb{X}_n|$ is the number of sample points contained in the ball $\mathcal{B}(X_j, h)$, and $\bar{X}_j = \frac{1}{N_j}\sum_{i\neq j} X_i \mathbb{1}_{\mathcal{B}(X_j,h)}(X_i)$ is their barycenter. Set $\hat{T}_j(h)$ to be the linear space spanned by the

d eigenvectors associated with the d largest eigenvalues of $\hat{\Sigma}_j(h)$. Computing a basis of $\hat{T}_j(h)$ can be performed naively using a singular value decomposition of the full matrix $\hat{\Sigma}_j(h)$, although fast PCA algorithms [30] may lessen the computational dependence on the ambient dimension. We also denote by $\texttt{TSE}(., h)$ the function which maps any vector of points to the vector of their estimated tangent spaces, with

$$\hat{T}_j(h) = \texttt{TSE}(\mathbb{X}_n, h)_j.$$

Proposition 14. *Set* $h = \left(c_{d,f_{min},f_{max}} \frac{\log n}{n-1}\right)^{1/d}$. *Then, for all* $Q \in \mathcal{G}_{D,d,f_{min},f_{max},\rho}$, *and* n *large enough, we have*

$$\max_{1 \leq j \leq n} \angle(T_{X_j} M, \hat{T}_j(h)) \leq C_{d,f_{min},f_{max}} \frac{h}{\rho},$$

with probability larger than $1 - 6\left(\frac{1}{n-1}\right)^{\frac{2}{d}}$.

Furthermore, it is shown in Lemma 30, based on the results of [10], that for $c_{d,f_{min},f_{max}}$ large enough, $\mathbb{X}_n$ is $c_{d,f_{min},f_{max}}h$-dense in M with probability larger than $1 - \left(\frac{1}{n}\right)^{2/d}$. Since $\mathbb{X}_n$ may not be sparse at the scale $c_{d,f_{min},f_{max}}h$, and for the stability reasons described in Section 3, we sparsify it with the farthest point sampling algorithm (Section 3.3) with scale parameter $\varepsilon = c_{d,f_{min},f_{max}}h$. Let $\mathbb{Y}_n$ denote the output of the algorithm.

Corollary 15. *With the above notation, for* n *large enough, with probability at least* $1 - 7\left(\frac{1}{n-1}\right)^{2/d}$,

- $\max_{X_j \in \mathbb{Y}_n} \angle(T_{X_j} M, \hat{T}_j(h)) \leq C_{d,f_{min},f_{max}} \frac{\varepsilon}{\rho}$,
- $\max_{x \in M} \mathrm{d}(x, \mathbb{Y}_n) \leq 2\varepsilon$.
- $\mathbb{Y}_n$ *is* ε*-sparse,*

In other words, the previous result shows that $\mathbb{Y}_n$ satisfies the assumptions of Theorem 13. We may then define $\hat{M}_{\texttt{TDC}}$ to be the Tangential Delaunay Complex computed on $\mathbb{Y}_n$ and the collection of estimated tangent spaces $\texttt{TSE}(\mathbb{X}_n, h)_{\mathbb{Y}_n}$, that is elements of $\texttt{TSE}(\mathbb{X}_n, h)$ corresponding to elements of $\mathbb{Y}_n$, where h is the bandwidth defined in Proposition 14.

Definition 16. *With the above notation, define* $\hat{M}_{\texttt{TDC}} = \mathrm{Del}^{\omega_*}\left(\mathbb{Y}_n, \texttt{TSE}(\mathbb{X}_n, h)_{\mathbb{Y}_n}\right)$.

Combining Theorem 13 and Corollary 15, it is clear that $\hat{M}_{\texttt{TDC}}$ satisfies Theorem 6.

5.2 Clutter Noise Case

We now address the denoising problem, following ideas from [23]. To distinguish whether X_j is an outlier or belongs to M, we notice again that points drawn from M approximately lie on a low dimensional structure. On the other hand, the neighborhood points of an outlier drawn far away from M should typically be distributed in an isotropic way. Let $k_1, k_2, h > 0$, $x \in \mathbb{R}^D$ and $T \subset \mathbb{R}^D$ a d-dimensional linear subspace. The *slab* at x in the direction T is the set $S(x, T, h) = \{x\} \oplus \mathcal{B}_T(0, k_1 h) \oplus \mathcal{B}_{T^\perp}\left(0, k_2 h^2\right) \subset \mathbb{R}^D$, where $\oplus$ denotes the Minkovski sum, and $\mathcal{B}_T, \mathcal{B}_{T^\perp}$ are the Euclidean balls in T and $T^\perp$ respectively.

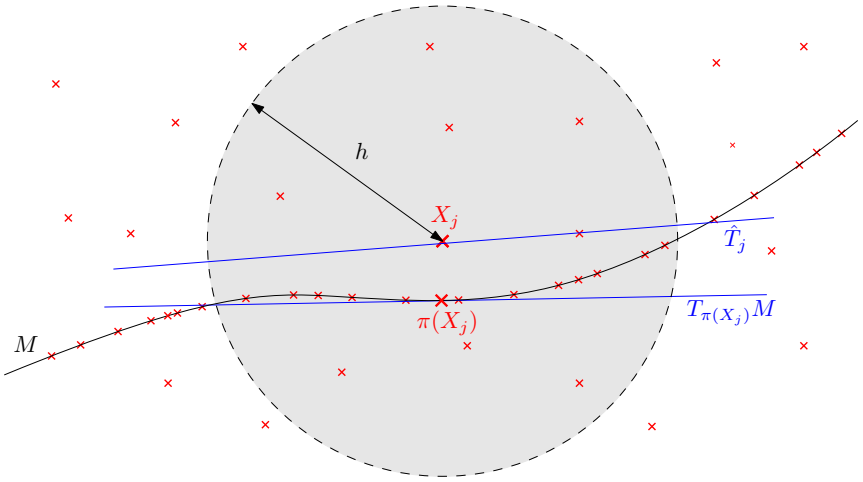


Figure 4: Local PCA at an outlier point $X_j \in \mathbb{X}_n$.

Using the notation of Section 2.1, for all $P \in \mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}$, Figure 5 indicates that

$$\begin{array}{lll} P\left(S(x, T_{\pi(x)}M, h)\right) \asymp h^d & \text{if} & \mathrm{d}(x, M) \leq h^2, \\ P\left(S(x, T_{\pi(x)}, h)\right) \asymp h^{2D-d} & \text{if} & \mathrm{d}(x, M) > h^2, \end{array}$$

as h goes to 0, for k_1 and k_2 small enough. Since $h^{2D-d} \ll h^d$, the measure of the slabs $P\left(S(x, T, h)\right)$ clearly is discriminatory for denoising, provided that tangent spaces are known.

Based on this intuition, we define the elementary step of our denoising procedure as the following map $\mathtt{SD}_t(.,.,h)$, that sends a vector $P = (p_1, \ldots, p_r) \subset \mathbb{R}^D$ and a corresponding vector of (estimated) tangent spaces $T_P = (T_1, \ldots, T_r)$ onto a subvector of P according to the rule

$$p_j \in \mathtt{SD}_t(P, T_P, h) \quad \Leftrightarrow \quad |S(p_j, T_j, h) \cap P| \geq t \log(n-1),$$

where t is a threshold to be fixed. This procedure relies on counting how many sample points lie in the slabs of direction the estimated tangent spaces (see Figure 5).

Since tangent spaces are unknown, the following Lemma gives some insight on the relation between the accuracy of the tangent space estimation and the denoising performance that can be reached.

Lemma 17. *Let $K > 0$ be fixed. There exists constants $k_1(K, d)$ and $k_2(d, D, \rho)$ such that, for every $h \leq h_+ \wedge 1$ and x we have*

$$\begin{array}{rcl} \mathrm{d}(x, M) \geq h/\sqrt{2} & \Rightarrow & S(x, T, h) \cap M = \emptyset, \\ \mathrm{d}(x, M) \geq h^2/\rho \quad \text{and} \quad \angle\left(T_{\pi(x)}M, T\right) \leq Kh/\rho & \Rightarrow & S(x, T, h) \cap M = \emptyset. \end{array}$$

Furthermore, if x and y are in M, then there exists a constant k_3 such that

$$\|x - y\| \leq k_3 h \quad \Rightarrow \quad y \in S(x, T_{\pi(x)}M, h)(x).$$

Possible values for k_1 and k_2 are, respectively, $\frac{3}{4d+8K\sqrt{d}}$ and $\frac{1}{4\sqrt{D-d}(\rho \vee 1)}$. Then k_3 may be chosen as $\frac{k_2\rho}{2K} \wedge \frac{k_1}{2} \wedge \sqrt{\rho k_1} \wedge \sqrt{\rho k_2}$.

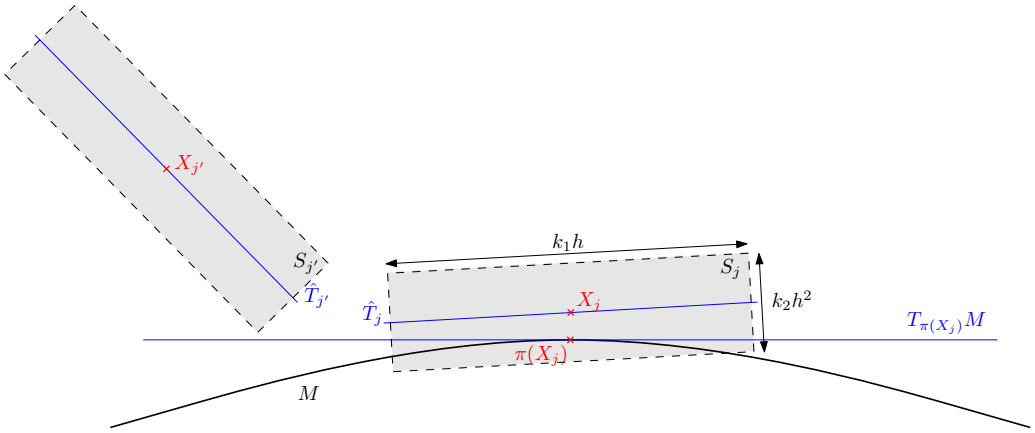


Figure 5: The slab S_j is centered at X_j and has size $k_1 h$ in the d directions spanned by $\hat{T}_j$, and size $k_2 h^2$ in the $D-d$ directions normal to $\hat{T}_j$.

The proof of Lemma 17, mentioned in [23], follows from elementary geometry, combined with the definition of the reach and Proposition 26.

Roughly, Lemma 17 states that the denoising performance is of order the square of the tangent space precision, hence will be closely related to the performance of the tangent space estimation procedure TSE. Unfortunately, a direct application of TSE to the corrupted sample $\mathbb{X}_n$ leads to slightly worse precision bounds, in terms of angle deviation. Typically, the angle deviation would be of order $n^{-1/(d+1)}$. However, this precision is enough to remove outliers points which are at distance at least $n^{-2/(d+1)}$ from M. Then running our TSE on this refined sample $\mathtt{SD}_t(\mathbb{X}_n, \mathtt{TSE}(\mathbb{X}_n), n^{-1/(d+1)})$ leads to better angle deviation rates, hence better denoising performance, and so on.

Let us now introduce the iterative denoising procedure in a more formal way. We choose the initial bandwidth $h_0 = \left(c_{d,f_{min},f_{max},\rho} \frac{\log n}{\beta(n-1)}\right)^{\gamma_0}$, with $\gamma_0 = 1/(d+1)$, and define the first set $\mathbb{X}^{(-1)} = \mathbb{X}_n$ as the whole sample. We then proceed recursively, setting $h_{k+1} = \left(c_{d,f_{min},f_{max},\rho} \frac{\log n}{\beta(n-1)}\right)^{\gamma_{k+1}}$, with γ_{k+1} satisfying $\gamma_{k+1} = (2\gamma_k + 1)/(d+2)$, and defining

$$\mathbb{X}^{(k+1)} = \mathtt{SD}_t(\mathbb{X}^{(k)}, \mathtt{TSE}(\mathbb{X}^{(k)}, h_{k+1}), h_{k+1}).$$

In other words, at step $k+1$ we use a smaller bandwidth h_{k+1} in the tangent space estimation procedure TSE. Then we use this better estimation of tangent spaces to run the elementary denoising step SD. The performance of this procedure is guaranteed by the next Proposition. With a slight abuse of notation, if X_j is in $\mathbb{X}^{(k)}$, $\mathtt{TSE}(\mathbb{X}^{(k)}, h)_j$ will denote the corresponding tangent space of $\mathtt{TSE}(\mathbb{X}^{(k)}, h)$.

Proposition 18. *In the clutter noise model, for t, $c_{d,f_{min},f_{max}}$ and n large enough, k_1 and k_2 small enough, the following properties hold with probability larger than $1 - 6\left(\frac{1}{n-1}\right)^{2/d}$.*

***Initialization**:*

- *For all $X_j \in \mathbb{X}^{(-1)}$ such that $\mathrm{d}(X_j, M) \leq h_0/\sqrt{2}$,*

$$\angle(\mathtt{TSE}(\mathbb{X}^{(-1)}, h_0)_j, T_{\pi(X_j)}M) \leq C_{d,f_{min},f_{max}} h_0/\rho.$$

- *For every* $X_j \in M \cap \mathbb{X}^{(-1)}$, $X_j \in \mathbb{X}^{(0)}$.
- *For every* $X_j \in \mathbb{X}^{(-1)}$, *if* $\mathrm{d}(X_j, M) > h_0^2/\rho$, *then* $X_j \notin \mathbb{X}^{(0)}$.

Iteration:

- *For all* $X_j \in \mathbb{X}^{(k)}$ *such that* $\mathrm{d}(X_j, M) \leq h_{k+1}/\sqrt{2}$,

$$\angle(\mathtt{TSE}(\mathbb{X}^{(k)}, h_{k+1})_j, T_{\pi(X_j)}M) \leq C_{d,f_{min},f_{max}} h_{k+1}/\rho.$$

- *For every* $X_j \in M \cap \mathbb{X}^{(k)}$, $X_j \in \mathbb{X}^{(k+1)}$.
- *For every* $X_j \in \mathbb{X}^{(k)}$, *if* $\mathrm{d}(X_j, M) > h_{k+1}^2/\rho$, *then* $X_j \notin \mathbb{X}^{(k+1)}$.

This result is threefold. Not only can we distinguish data and outliers within a decreasing sequence of offsets of radii h_k^2/ρ around M, but we can also ensure that no points of M is removed during the process, with high probability. Moreover, it also provides a convergence rate for the estimated tangent spaces $\mathtt{TSE}(\mathbb{X}_k, h_{k+1})$.

Now fix a precision level δ. An elementary calculation shows that if k is larger than $(\log(1/\delta) - \log(d(d+1))/(\log(d+2) - \log(2))$, then $1/d > \gamma_k \geq 1/d - \delta$. Let us then define k_δ as the smallest integer satisfying $\gamma_k \geq 1/d - \delta$, and denote by $\mathbb{Y}_n^\delta$ the output of the farthest point sampling algorithm applied to $\mathbb{X}^{(k_\delta)}$ with parameter $\varepsilon = c_{d,f_{min}f_{max}} h_{k_\delta}$ (for $c_{d,f_{min}f_{max}}$ large enough). Define also T^δ as the restriction of $\mathtt{TSE}(\mathbb{X}^{(k_\delta)}, h_{k_\delta})$ to the elements of $\mathbb{Y}_n^\delta$.

According to Proposition 18, the denoising procedure removes no data point on M with high probability. In other words, $\mathbb{X}^{(k_\delta)} \cap M = \mathbb{X}_n \cap M$, and as a consequence, $\max_{x \in M} \mathrm{d}(x, \mathbb{X}^{(k_\delta)}) \leq c_{d,f_{min}} \left(\frac{\log n}{\beta n}\right)^{1/d} \ll h_{k_\delta}$ with high probability, using the same result of [10] as in the noise-free case (see Lemma 30 at Section C).

Corollary 19. *With the above notation, for* n *large enough, with probability larger than* $1 - 7\left(\frac{1}{n-1}\right)^{2/d}$,

- $\max_{X_j \in \mathbb{Y}_n^\delta} \mathrm{d}(X_j, M) \leq \varepsilon^2$,
- $\mathbb{Y}_n^\delta$ *is* ε*-sparse,*
- $\max_{X_j \in \mathbb{Y}_n^\delta} \angle(T_j^\delta, T_{\pi(X_j)}M) \leq C_{d,f_{min},f_{max},\rho}\varepsilon$,
- $\max_{x \in M} \mathrm{d}(x, \mathbb{Y}_n^\delta) \leq 2\varepsilon$.

We are now able to define the estimator $\hat{M}_{\mathtt{TDC}\delta}$.

Definition 20. *With the above notation, define* $\hat{M}_{\mathtt{TDC}\delta} = \mathrm{Del}^{\omega_*}\left(\mathbb{Y}_n^\delta, T^\delta\right)$.

Combining Theorem 13 and Corollary 19, it is clear that $\hat{M}_{\mathtt{TDC}\delta}$ satisfies Theorem 7.

Finally, we turn to the asymptotic estimator $\hat{M}_{\mathtt{TDC}+}$. Set $h_\infty = \left(c_{d,f_{min},f_{max},\rho} \frac{\log n}{\beta(n-1)}\right)^{1/d}$, and let $\hat{k}$ denote the smallest integer such that $\min\{\mathrm{d}(X_j, M) | \mathrm{d}(X_j, M) > h_\infty^2/\rho\} > h_{\hat{k}}^2/\rho$. Since $\mathbb{X}_n$ is a (random) finite set, we can always find such a random integer $\hat{k}$ that provides a sufficient number of iterations to obtain the asymptotic denoising rate. For this random iteration $\hat{k}$, we can state the following result.

Proposition 21. *Under the Assumptions of Proposition 19, for every $X_j \in X^{(\hat{k}+1)}$, we have*

$$\angle(\texttt{TSE}(\mathbb{X}^{(\hat{k}+1)}, h_\infty)_j, T_{\pi(X_j)}M) \leq C_{d,f_{min},f_{max}} h_\infty/\rho.$$

As before, taking $\mathbb{Y}_n^+$ as the result of the farthest point sampling algorithm based on $\mathbb{X}^{(\hat{k}+1)}$, and T^+ the vector of tangent spaces $\texttt{TSE}(\mathbb{X}^{(\hat{k}+1)}, h_\infty)_j$ such that $\mathbb{X}_j^{(\hat{k}+1)} \in \mathbb{Y}_n^+$, we can construct our last estimator.

Definition 22. *With the above notation, define* $\hat{M}_{\texttt{TDC+}} = \mathrm{Del}^{\omega_*}(\mathbb{Y}_n^+, T^+)$.

In turn, Proposition 21 implies that $\hat{M}_{\texttt{TDC+}}$ satisfies Theorem 8.

6 Conclusion

In this work, we gave results on explicit manifold reconstruction with simplicial complexes. We built estimators $\hat{M}_{\texttt{TDC}}$, $\hat{M}_{\texttt{TDC}\delta}$ and $\hat{M}_{\texttt{TDC+}}$ in two statistical models. We proved minimax rates of convergence for the Hausdorff distance and consistency results for ambient isotopic reconstruction. Since $\hat{M}_{\texttt{TDC}}$ is minimax optimal and uses the Tangential Delaunay Complex of [4], the latter is proved to be optimal. Moreover, rates of [23] are proved to be achievable with simplicial complexes that are computable using existing algorithms. To prove the stability of the Tangential Delaunay Complex, a generic interpolation result was derived. In the process, a tangent space estimation procedure and a denoising method both based on local PCA were studied.

In the model with outliers, the proposed reconstruction method achieves a rate of convergence that can be as close as desired to the minimax rate of convergence, depending on the number of iterations of our denoising procedure. Though this procedure seems to be well adapted to our reconstuction scheme which is based on tangent spaces estimation, we believe that it could be of interest in the context of other applications. Also, further investigation may be carried out to compare this denoising procedure to the existing ones (see, e.g., [9], [18]).

The effective construction of $\hat{M}_{TDC\delta}$ can be performed using existing algorithms. Namely, Tangential Delaunay Complex, farthest point sampling, local PCA and point-to-linear subspace distance computation for slab counting. A crude upper bound on the time complexity of a naive step-by-step implementation is

$$O\left(nD\left[2^{O(d^2)} + \log(1/\delta)D(D+n)\right]\right),$$

since the precision δ necessitates no more than $\log(1/\delta)$ iterations of the denoising procedure. It is likely that better complexity bounds may be obtained using more refined algorithms, such as fast PCA (see, e.g., [30]). An interesting development would be to investigate this suggested precision/complexity tradeoff, as done in [3] for community detection in graphs for instance.

Even though Theorem 10 is applied to manifold estimation, the authors believe it may be applied in various settings. Beyond its statement, the way that it is used is quite general. When intermediate objects (here, tangent spaces) are used in a procedure, this kind of proxy method can provide extensions of existing results to the case where these objects are only approximated.

As local PCA is performed throughout the paper, the knowledge of the bandwidth h is needed for actual implementation. In practice its choice is a difficult question and adaptive selection of h remains to be considered.

In the process, we derived rates of convergence for tangent space estimation. The optimality of the method will be the object of a future paper.

Acknowledgements

The authors would like to thank Jean-Daniel Boissonnat , Frédéric Chazal, Pascal Massart, and Steve Oudot for their insight and the interest they brought to this work.

This work was supported by ANR project TopData ANR-13-BS01-0008 and by the Advanced Grant of the European Research Council GUDHI (Geometric Understanding in Higher Dimensions). E. Aamari was supported by the Conseil régional d'Île-de-France under a doctoral allowance of its program Réseau de Recherche Doctoral en Mathématiques de l'Île-de-France (RDM-IdF).

References

[1] Stephanie B. Alexander and Richard L. Bishop. Gauss equation and injectivity radii for subspaces in spaces of curvature bounded above. Geom. Dedicata, 117:65–84, 2006.

[2] E. Arias-Castro, G. Lerman, and T. Zhang. Spectral Clustering Based on Local PCA. ArXiv e-prints, January 2013.

[3] Ery Arias-Castro and Nicolas Verzelen. Community detection in dense random networks. Ann. Statist., 42(3):940–969, 2014.

[4] Jean-Daniel Boissonnat and Arijit Ghosh. Manifold reconstruction using tangential Delaunay complexes. Discrete Comput. Geom., 51(1):221–267, 2014.

[5] Jean-Daniel Boissonnat, Leonidas J. Guibas, and Steve Y. Oudot. Manifold reconstruction in arbitrary dimensions using witness complexes. Discrete Comput. Geom., 42(1):37–70, 2009.

[6] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. ESAIM Probab. Stat., 9:323–375, 2005.

[7] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[8] Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. C. R. Math. Acad. Sci. Paris, 334(6):495–500, 2002.

[9] Mickaël Buchet, Tamal K. Dey, Jiayuan Wang, and Yusu Wang. Declutter and resample: Towards parameter free denoising. CoRR, abs/1511.05479, 2015.

[10] F. Chazal, M. Glisse, C. Labruère, and B. Michel. Optimal rates of convergence for persistence diagrams in Topological Data Analysis. ArXiv e-prints, May 2013.

[11] Frédéric Chazal, David Cohen-Steiner, and André Lieutier. A sampling theory for compact sets in Euclidean space. In Computational geometry (SCG'06), pages 319–326. ACM, New York, 2006.

[12] Siu-Wing Cheng, Tamal K. Dey, and Edgar A. Ramos. Manifold reconstruction from point samples. In Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 1018–1027. ACM, New York, 2005.

[13] Kenneth L Clarkson. Building triangulations using ε-nets. In Proceedings of the thirty-eighth annual ACM symposium on Theory of computing, pages 326–335. ACM, 2006.

[14] Antonio Cuevas and Alberto Rodríguez-Casal. On boundary estimation. Adv. in Appl. Probab., 36(2):340–354, 2004.

[15] Giuseppe De Marco, Gianluca Gorni, and Gaetano Zampieri. Global inversion of functions: an introduction. NoDEA Nonlinear Differential Equations Appl., 1(3):229–248, 1994.

[16] Tamal K. Dey. Curve and surface reconstruction: algorithms with mathematical analysis, volume 23 of Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, 2007.

[17] Tamal K. Dey and Kuiyu Li. Topology from data via geodesic complexes. Technical report OSU-CISRC-3/09-TR05.

[18] D. L. Donoho. De-noising by soft-thresholding. IEEE Trans. Inf. Theor., 41(3):613–627, May 1995.

[19] Lutz Dümbgen and Günther Walther. Rates of convergence for random approximations of convex sets. Adv. in Appl. Probab., 28(2):384–393, 1996.

[20] Ramsay Dyer, Gert Vegter, and Mathijs Wintraecken. Riemannian Simplices and Triangulations. In Lars Arge and János Pach, editors, 31st International Symposium on Computational Geometry (SoCG 2015), volume 34 of Leibniz International Proceedings in Informatics (LIPIcs), pages 255–269, Dagstuhl, Germany, 2015. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[21] Herbert Federer. Curvature measures. Trans. Amer. Math. Soc., 93:418–491, 1959.

[22] Herbert Federer. Geometric measure theory. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York, 1969.

[23] Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. Ann. Statist., 40(2):941–963, 2012.

[24] Gene H. Golub and Charles F. Van Loan. Matrix computations. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[25] Arlene K. H. Kim and Harrison H. Zhou. Tight minimax rates for manifold estimation under Hausdorff loss. Electron. J. Stat., 9(1):1562–1582, 2015.

[26] L. LeCam. Convergence of estimates under dimensionality restrictions. Ann. Statist., 1:38–53, 1973.

[27] Michel Ledoux and Michel Talagrand. Probability in Banach spaces, volume 23 of Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

[28] E. Mammen and A. B. Tsybakov. Asymptotical minimax recovery of sets with smooth boundaries. Ann. Statist., 23(2):502–524, 1995.

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

[30] Alok Sharma and Kuldip K Paliwal. Fast principal component analysis using fixed-point algorithm. Pattern Recognition Letters, 28(10):1151–1155, 2007.

[31] Sara van de Geer. Empirical processes in M-estimation. Cambridge University Press, 2000.

A Interpolation Theorem

This section is devoted to prove the interpolation results of Section 4.1. For sake of completeness, let us state a stability result for the reach with respect to $\mathcal{C}^2$ diffeomorphisms.

Lemma 23 (Theorem 4.19 in [21])**.** *Let $A \subset \mathbb{R}^D$ with $\mathrm{reach}(A) \geq \rho > 0$ and $\Phi : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ is a $\mathcal{C}^1$-diffeomorphism such that Φ,Φ^{-1}, and $\mathrm{d}\Phi$ are Lipschitz with Lipschitz constants K,N and R respectively, then*

$$\mathrm{reach}(\Phi(A)) \geq \frac{1}{(K\rho^{-1} + R)N^2}.$$

Writing $\phi_\ell(\cdot) = \phi(\cdot/\ell)$, we recall that $\psi_j(a) = (R_j - I_D)(a - \pi(p_j)) + (p_j - \pi(p_j))$ and

$$\Phi(a) = a + \sum_{j=1}^{q} \phi_\ell\left(a - \pi(p_j)\right) \psi_j(a). \tag{1}$$

Let us denote $b_1 = \sup_x \|\mathrm{d}_x\phi\|$, $b_2 = \sup_x \left\|\mathrm{d}^2_x\phi\right\|_{\mathrm{op}}$, and write $C_1 = 1 + b_1$, $C_2 = b_2 + 2b_1$. Straightforward computation yields $C_1 \leq 7/2$ and $C_2 \leq 28$.

Proof of Lemma 11. First notice that the sum appearing in (1) consists of at most one term. Indeed, since $\phi \equiv 0$ outside $\mathcal{B}(0,1)$, if $\phi_\ell\left(a - \pi(p_j)\right) \neq 0$ for some $j \in \{1, \ldots, q\}$, then $\|a - \pi(p_j)\| \leq \ell$. Consequently, for all $i \neq j$,

$$\begin{aligned}
\|a - \pi(p_i)\| &\geq \|p_j - p_i\| - \|p_j - \pi(p_j)\| - \|\pi(p_j) - a\| - \|\pi(p_i) - p_i\| \\
&\geq \delta - \eta - \ell - \eta \\
&\geq \delta - 2\ell \geq \ell,
\end{aligned}$$

where we used that $6\eta \leq \ell \leq \delta/3$. Therefore, $\phi_\ell\left(a - \pi(p_i)\right) = 0$ for all $i \neq j$. In other words, if a p_j actually appears in $\Phi(a)$ then the others do not.

Global diffeomorphism: As the sum in (1) is at most composed of one term, chain rule yields

$$\begin{aligned}
\left\|\mathrm{d}_a\Phi - \mathrm{Id}\right\|_{\mathrm{op}} &= \max_{1\leq j\leq q} \left\|\mathrm{d}_a\left[\phi_\ell\left(a - \pi(p_j)\right)\psi_j(a)\right]\right\|_{\mathrm{op}} \\
&= \max_{1\leq j\leq q} \left\| \psi_j(a) \left.\frac{\mathrm{d}_b\phi}{\ell}\right|_{b=\frac{a-\pi(p_j)}{\ell}} + \phi_\ell\left(a - \pi(p_j)\right)\left(R_j - I_D\right) \right\|_{\mathrm{op}} \\
&\leq \big(b_1 + 1\big)\theta + b_1\frac{\eta}{\ell} < 1,
\end{aligned}$$

where the last line follows from $b_1 \leq 5/2$, $6\eta \leq \ell$ and $\theta \leq \pi/64$.

Therefore, $\mathrm{d}_a\Phi$ is invertible for all $a \in \mathbb{R}^D$, and $(\mathrm{d}_a\Phi)^{-1} = \sum_{i=0}^{\infty} (I_D - \mathrm{d}_a\Phi)^i$. Φ is a local diffeomorphism according to the local inverse function theorem. Moreover, $\|\Phi(a)\| \to \infty$ as $\|a\| \to \infty$, so that Φ is a global $\mathcal{C}^\infty$-diffeomorphism by Hadamard-Cacciopoli theorem [15].

Differentials estimates: (i) First order: From the estimates above,

$$\|\mathrm{d}_a\Phi\|_{\mathrm{op}} \leq \|I_D\|_{op} + \|\mathrm{d}_a\Phi - I_D\|_{op} \leq 1 + \big(b_1 + 1\big)\theta + b_1\frac{\eta}{\ell}.$$

(ii) Inverse: Write for all $a \in \mathbb{R}^D$,

$$\begin{aligned}\left\|\mathrm{d}_{\Phi(a)}\Phi^{-1}\right\|_{\mathrm{op}} = \left\|(\mathrm{d}_a\Phi)^{-1}\right\|_{\mathrm{op}} &= \left\|\sum_{i=0}^{\infty}(I_D - \mathrm{d}_a\Phi)^i\right\|_{\mathrm{op}} \\ &\leq \frac{1}{1 - \|\mathrm{Id} - \mathrm{d}_a\Phi\|_{\mathrm{op}}} \leq \frac{1}{1 - \big(b_1+1\big)\theta - b_1\frac{\eta}{\ell}},\end{aligned}$$

where the first inequality holds since $\|\mathrm{d}_a\Phi - \mathrm{Id}\|_{\mathrm{op}} < 1$, and $\|\cdot\|_{\mathrm{op}}$ is sub-multiplicative.
(iii) Second order: Again, since the sum (1) includes at most one term,

$$\begin{aligned}\left\|\mathrm{d}_a^2\Phi\right\|_{\mathrm{op}} &= \max_{1\leq j\leq q}\left\|\mathrm{d}_a^2\left[\phi_\ell\left(a - \pi(p_j)\right)\psi_j(a)\right]\right\|_{\mathrm{op}} \\ &\leq \max_{1\leq j\leq q}\left\{\frac{\left\|\mathrm{d}^2\phi\right\|_{\mathrm{op}}}{\ell^2}\|\psi_j(a)\| + 2\frac{\|\mathrm{d}\phi\|_{\mathrm{op}}}{\ell}\|R_j - I_D\|_{\mathrm{op}}\right\} \\ &\leq b_2\frac{\eta}{\ell^2} + \left(b_2 + 2b_1\right)\frac{\theta}{\ell}.\end{aligned}$$

□

Proof of Theorem 10. Set $\ell = \delta/3$ and $M' = \Phi(M)$.

- *Interpolation:* For all j, $p_j = \Phi(\pi(p_j)) \in M'$ by construction since $\phi_\ell(0) = 1$.
- *Tangent spaces:* Since $\mathrm{d}_x\phi_l|_{x=0} = 0$, for all $j \in \big\{1, \ldots, q\big\}$, $\mathrm{d}_a\Phi|_{a=\pi(p_j)} = R_j$. Thus,

$$\begin{aligned}T_{p_j}M' &= T_{\Phi(\pi(p_j))}\Phi(M) \\ &= \mathrm{d}_a\Phi|_{a=\pi(p_j)}\left(T_{\pi(p_j)}M\right) \\ &= R_j\left(T_{\pi(p_j)}M\right) = T_j,\end{aligned}$$

  by definition of R_j.
- *Proximity to M:* It follows from the correspondence

$$\begin{aligned}\|\Phi(a) - a\| &\leq \sup_{a\in\mathbb{R}^D}\max_{1\leq j\leq q}\phi_\ell\left(a - \pi(p_j)\right)\|\psi_j(a)\| \\ &\leq \ell\theta + \eta \leq \delta\theta + \eta.\end{aligned}$$

- *Isotopy:* Consider the continuous family of maps

$$\Phi_{(t)}(a) = a + t\left(\sum_{j=1}^{q}\phi_\ell\left(a - \pi(p_j)\right)\psi_j(a)\right),$$

for $0 \leq t \leq 1$. Since $\Phi_{(t)} - \mathrm{Id} = t\big(\Phi - \mathrm{Id}\big)$, the arguments above show that $\Phi_{(t)}$ is a global diffeomorphism of $\mathbb{R}^D$ for all $t \in [0,1]$. Moreover $\Phi_{(0)} = \mathrm{Id}$, and $\Phi_{(1)} = \Phi$. Thus, $M = \Phi_{(0)}(M)$ and $M' = \Phi_{(1)}(M)$ are ambient isotopic.

- *Reach lower bound:* The differentials estimates of order 1 and 2 of Φ translate into estimates on Lipschitz constants of Φ,Φ^{-1} and $\mathrm{d}\Phi$. Applying Lemma 23 leads to

$$\mathrm{reach}\,(M') \geq \frac{\big(1 - C_1\left(\frac{\eta}{\ell} + \theta\right)\big)^2}{\frac{1 + C_1\left(\frac{\eta}{\ell} + \theta\right)}{\rho} + C_2\left(\frac{\eta}{\ell^2} + \frac{\theta}{\ell}\right)} = \rho \cdot \frac{\big(1 - C_1\left(\frac{\eta}{\ell} + \theta\right)\big)^2}{1 + C_1\left(\frac{\eta}{\ell} + \theta\right) + C_2\left(\frac{\eta}{\ell^2} + \frac{\theta}{\ell}\right)\rho}.$$

  The desired lower bound follows by replacing ℓ by its value $\delta/3$, and setting $c_1 = 3C_1 \leq 21/2 \leq 11$ and $c_2 = 3^2 C_2 \leq 252$.

□

B Some Geometric Properties under Reach Regularity Condition

In this section we state intermediate results that connect Euclidean an geodesic quantities under reach regularity condition. We begin with a result connecting reach and principal curvatures.

Proposition 24 (Proposition 2.1 in [17])**.** *For all $x \in M$, writing II_x for the second fundamental form of M at x, for all unitary $w \in T_xM$, we have $\|II_x(w,w)\| \leq 1/\rho$.*

For all $x \in M$ and $v \in T_xM$, let us denote by $\exp_x(v)$ the exponential map at x of direction v. According to the following Proposition, this exponential map turns out to be a diffeomorphism on balls of radius at most $\pi\rho$.

Proposition 25 (Corollary 1.4 in [1])**.** *The injectivity radius of M is at least $\pi\rho$.*

Denoting by $\mathrm{d}_M(\cdot,\cdot)$ the geodesic distance on M, we are in position to connect geodesic and Euclidean distance. In what follows, we fix the constant $\alpha = 1 + \frac{1}{4\sqrt{2}}$.

Proposition 26. *For all $x, y \in M$ such that $\|x - y\| \leq \rho/4$,*

$$\|x - y\| \leq \mathrm{d}_M(x,y) \leq \alpha\,\|x - y\|\,.$$

Moreover, writing $y = \exp_x(rv)$ for $v \in T_xM$ with $\|v\| = 1$ and $r \leq \rho/4$,

$$y = x + rv + R(r,v)$$

with $\|R(r,v\| \leq \frac{r^2}{2\rho}$. As a consequence,

$$\|x - y\| \leq \mathrm{d}_M(x,y) \leq \|x - y\| + \frac{\alpha^2\,\|x - y\|^2}{2\rho}.$$

Proof of Proposition 26. The first statement is a direct consequence of Proposition 6.3 in [29]. Let us define $u(t) = \exp_x(tv) - \exp_x(0) - tv$ and $w(t) = \exp_x(tv)$ for all $0 \leq t \leq r$. It is clear that $u(0) = 0$ and $u'(0) = 0$. Moreover, $\|u''(t)\| = \left\|II_{w(t)}\left(w'(t), w'(t)\right)\right\| \leq 1/\rho$. Therefore, a Taylor expansion at order two gives $\|R(r,v)\| = u(r) \leq r^2/(2\rho)$. Applying the first statement of the proposition gives $r \leq \alpha \|x - y\|$. Therefore,

$$\|x - y\| \leq \mathrm{d}_M(x,y) \leq r \leq \|x - y\| + \|R(r,v)\| \leq \|x - y\| + \frac{\alpha^2 \|x - y\|^2}{2\rho}.$$

□

The next proposition gives bounds on the volume form expressed in polar coordinates in a neighborhood of points of M.

Proposition 27. *Let $x \in M$ be fixed. Denote by $J(r,v)$ the Jacobian of the volume form expressed in polar coordinates around x, for $r \leq \frac{\rho}{4}$ and v a unit vector in T_xM. In other words, if $y = \exp_x(rv)$, $dV_y = J(r,v)drdv$. Then*

$$c_d r^{d-1} \leq J(r,v) \leq C_d r^{d-1},$$

where $c_d = 2^{-d}$ and $C_d = 2^d$.

Proof of Proposition 27. Denoting $A_{r,v} = \mathrm{d}_{rv}\exp_x$, the Area Formula (see, *e.g.*, Section 3.2.5 in [22]) asserts that $J(r,v) = r^{d-1}\sqrt{\det\left(A_{r,v}^t A_{r,v}\right)}$. Note that from Proposition 2.1 in [17], the sectional curvatures in M are bounded by $|\kappa| \leq 2/\rho^2$. Therefore, the Rauch theorem (see, *e.g.*, Lemma 5 of [20]) states that

$$\left(1 - \frac{r^2}{3\rho^2}\right)\|w\| \leq \|A_{r,v}w\| \leq \left(1 + \frac{r^2}{\rho^2}\right)\|w\|,$$

for all $w \in T_xM$. As a consequence,

$$2^{-d} \leq \left(1 - \frac{r^2}{3\rho^2}\right)^d \leq \sqrt{\det\left(A_{r,v}^t A_{r,v}\right)} \leq \left(1 + \frac{r^2}{\rho^2}\right)^d \leq 2^d.$$

□

We will also need the following result giving an estimate of the size of projections of Euclidean balls onto M, when the center is close to M.

Proposition 28. *Let $x \in \mathbb{R}^D$ be such that $\mathrm{d}(x,M) = \Delta \leq h \leq \frac{\rho}{8}$, and let y denote $\pi(x)$. Then,*

$$\mathcal{B}\left(y, r_h^-\right) \cap M \subset \mathcal{B}(x,h) \cap M \subset \mathcal{B}\left(y, r_h^+\right) \cap M,$$

where $r_h^2 + \Delta^2 = h^2$, $r_h^- = \left(1 - \frac{\alpha^2\Delta}{\rho}\right) r_h$, and $r_h^+ = \left(1 + \frac{\alpha^2\Delta}{\rho}\right) r_h$.

Proof of Proposition 28. Let z be in $M \cap \mathcal{B}(x,h)$, and denote by δ the quantity $\|z-y\|$. Since $\delta \leq 2h$, according to proposition 26, we may write

$$\begin{aligned}\delta^2 &= \|z-x\|^2 - \|y-x\|^2 - 2\langle z-y, y-x\rangle \\ &\leq r_h^2 + \frac{\Delta\alpha^2}{\rho}\delta^2,\end{aligned}$$

hence $\delta \leq \left(1 + \frac{4\Delta\alpha^2}{\rho}\right) r_h$.

On the other hand, the same inequality ensures that, for any $z \in \mathcal{B}(y, \rho/4) \cap M$,

$$\|z-x\|^2 \leq \delta^2 + \Delta^2 + \frac{\alpha^2\Delta}{\rho}\delta^2.$$

Hence $\delta \leq \left(1 - \frac{\alpha^2\Delta}{\rho}\right) r_h$ ensures that $\|z-x\| \leq h$.

□

At last, the following consequence of Proposition 26 will be of particular use in the denoising procedure.

Proposition 29. *Let h and h_k be bandwidths satisfying $h_k^2/\rho \leq h \leq h_k \leq \rho/(12\alpha)$. Let x be such that $d(x,M) \leq h/\sqrt{2}$ and $\pi_M(x) = 0$, and let z be such that $d(z,x) \leq h$ and $d(z,M) \leq h_k^2/\rho$. Then*

$$\|z_\perp\| \leq \frac{10h_k^2}{\rho},$$

where $z_\perp$ denotes the projection of z onto $T_0M^\perp$.

Proof of Proposition 29. Let y denote $\pi_M(z)$. A triangle inequality yields $\|y\| \leq \|y-z\| + \|z-x\| + \|x\| \leq h_k^2/\rho + (1+1/\sqrt{2})h \leq \rho/4$. Proposition 26 ensures that $\|y_\perp\| \leq \alpha^2\|y\|^2/(2\rho)$, hence the result, noting that $\|z_\perp\| \leq \|y_\perp\| + h_k^2/\rho$. □

C Some Technical Properties of the Statistical Model

Lemma 30. *Let $P \in \mathcal{U}_M(f_{min}, f_{max})$. Then for all $p \in M$ and $r \leq \rho/4$,*

$$P(\mathcal{B}(p,r)) \geq a_d f_{min} r^d.$$

As a consequence, for n large enough and for all $Q \in \mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}$, with probability larger that $1 - \left(\frac{1}{n}\right)^{2/d}$,

$$\mathrm{d_H}(M, \mathbb{X}_n \cap M) \leq C_{d,f_{min}} \left(\frac{\log n}{\beta n}\right)^{1/d}.$$

Since the function $x \mapsto \log x/x$ is a decreasing function on $[3,\infty)$, the bound $\mathrm{d_H}(M, \mathbb{X}_n \cap M) \leq C_{d,f_{min}} \left(\frac{\log(n-1)}{\beta(n-1)}\right)^{1/d}$ also holds with probability at least $1 - \left(\frac{1}{n}\right)^{2/d}$.

Proof of Lemma 30. The first statement is a direct corollary of Proposition 27. Denoting $a = a_d f_{min}$, it corresponds to the (a,d)-standardness condition of [10] for small radii r. For all $k = 0,\dots,n$, conditionally on the event $\{|\mathbb{X}_n \cap M| = k\}$, $\mathbb{X}_n \cap M$ has the distribution of a k-sample of P. But from the previous point, P fulfils the (a,d)-standard assumption of [10] for $r \leq \rho/4$. Looking carefully at the proof of Theorem 3.3 in [10] shows that its conclusion still holds for measures satisfying the (a,d)-standard assumption for small radii only. Therefore, for $r \leq \rho/8$,

$$\mathbb{P}\left(\mathrm{d_H}(M, \mathbb{X}_n \cap M) > r || \mathbb{X}_n \cap M| = k\right) \leq \frac{4^d}{ar^d} \exp\left(-k\frac{a}{2^d} r^d\right).$$

Hence,

$$\begin{aligned}
\mathbb{P}\left(\mathrm{d_H}(M, \mathbb{X}_n \cap M) > r\right) &= \sum_{k=0}^{n} \mathbb{P}\left(\mathrm{d_H}(M, \mathbb{X}_n \cap M) > r | \, |\mathbb{X}_n \cap M| = k\right) \mathbb{P}(|\mathbb{X}_n \cap M| = k) \\
&\leq \sum_{k=0}^{n} \frac{4^d}{ar^d} \exp\left(-k\frac{a}{2^d} r^d\right) \binom{n}{k} \beta^k (1-\beta)^{n-k} \\
&= \frac{4^d}{ar^d} \left[1 - \beta\left(1 - \exp\left(-\frac{a}{2^d} r^d\right)\right)\right]^n \\
&\leq \frac{4^d}{ar^d} \exp\left[-n\beta\left(1 - \exp\left(-\frac{a}{2^d} r^d\right)\right)\right] \\
&\leq \frac{4^d}{ar^d} \exp\left[-n\beta \frac{a}{2^{d+1}} r^d\right],
\end{aligned}$$

whenever $r \leq \rho/8$ and $ar^d \leq 2^d$. Taking $r = C_{d,f_{min},\beta}\left(\frac{\log n}{n}\right)^{1/d}$ with $C^d_{d,f_{min},\beta} \frac{\beta a}{2^{d+1}} \geq 1 + 2/d$ yields the result. □

We are now able to prove Lemma 2.

Proof of Lemma 2. For $\varepsilon \leq \rho/4$, let $\mathrm{cv}_M(\varepsilon)$ be the geodesic covering number of M. $\mathrm{cv}_M(\varepsilon)$ is the minimal number k of geodesic balls $\mathcal{B}_M(x_1, \varepsilon), \dots, \mathcal{B}_M(x_k, \varepsilon)$ of radius ε needed to cover M. Notice that according to Lemma 30 and Proposition 26,

$$P\left(\mathcal{B}_M(x_k, \varepsilon)\right) \geq \frac{f_{min}\varepsilon^d}{c_d}.$$

Therefore, a packing argument (see Section B.1 of [10]) yields

$$\mathrm{cv}_M(\varepsilon) \leq \frac{2^d c_d}{f_{min}\varepsilon^d}.$$

Now, take $p, q \in M$ such that $\|p - q\| = \mathrm{diam}(M)$. Let γ be a minimal geodesic joining p and q. Since γ is minimal, it intersects each $\mathcal{B}_M(x_k, \varepsilon)$ at most once. Reorder the centers of the balls of the covering that intersect γ successively: $x_{(1)}, \dots, x_{(N)}$. Taking $\varepsilon = \rho/4$ and using triangle inequality,

we may write

$$\begin{aligned}\operatorname{diam}(M) &= \|p-q\| \\ &\leq \left\|p-x_{(1)}\right\| + \sum_{k=1}^{N-1} \left\|x_{(k)}-x_{(k+1)}\right\| + \left\|x_{(K)}-q\right\| \\ &\leq 2\varepsilon N \leq 2\varepsilon \mathrm{cv}_M(\varepsilon) \\ &\leq 2\frac{\rho}{4}\frac{2^d c_d}{f_{min}\left(\frac{\rho}{4}\right)^d} := \frac{C_d}{f_{min}\rho^{d-1}}.\end{aligned}$$

□

D Matrix Decomposition and Principal Angles

This section states results on the connections between matrix decomposition and principal angles between linear spans of eigenvectors. For symmetric matrices, we let $\lambda_i(\cdot)$ denote their i-th largest eigenvalue and $\lambda_{min}(\cdot)$ the smallest one.

Proposition 31. *Let $O \in \mathbb{R}^{D\times D}$, $B \in \mathbb{R}^{d\times d}$ be symmetric matrices such that*

$$O = \left(\begin{array}{c|c} B & 0 \\ \hline 0 & 0 \end{array}\right) + E,$$

$\lambda_{min}(B) \geq 1 - e_1$ and $\|E\|_{\mathcal{F}} \leq e_2$. Let T_0 (resp. T) be the vector space spanned by the first d vectors of the canonical basis, (resp. by the first d eigenvectors of O). If $e_1 + e_2 \leq 1/2$, then $\angle(T_0, T) \leq 2de_2$.

The proof of Proposition 31 relies on the application of Wielandt-Hoffmann Theorem, which is recalled below.

Theorem 32 (Wielandt-Hoffmann, Theorem 8.1.4 in [24])**.** *Let $A, E \in \mathbb{R}^{D\times D}$ be symmetric matrices. Then,*

$$\sum_{i=1}^{D} \left(\lambda_i(A+E) - \lambda_i(A)\right)^2 \leq \|E\|_{\mathcal{F}}^2.$$

Proof of Proposition 31. Denoting $\lambda_i = \lambda_i(O)$ and $b_i = \lambda_i(B)$, Theorem 32 yields

$$\sum_{i=1}^{d}(\lambda_i - b_i)^2 + \sum_{i=d+1}^{D} \lambda_i^2 \leq e_2^2.$$

Let $u_1, \ldots, u_d$ denote the first d eigenvectors of O (that span T). Moreover, denote by π_1 the orthogonal projection onto T_0 and by π_2 the orthogonal projection onto its orthogonal complement $T_0^{\perp}$. Then O takes the form $\tilde{\pi}_1 + E$, where $\tilde{\pi}_1$ has the same range as π_1. Let k be in $\{1, \ldots, d\}$, then, by definition

$$(\tilde{\pi}_1 + E)\, u_k = \lambda_k u_k.$$

Therefore

$$\pi_2 E u_k = \lambda_k \pi_2 u_k.$$

According to Theorem 32, $|\lambda_k - b_k| \leq e_1$, hence $\lambda_k \geq 1 - e_1 - e_2$. We deduce that

$$\|\pi_2 u_k\| \leq \frac{e_2}{1 - e_1 - e_2} \leq 2e_2.$$

As a consequence, for all $u \in T$, $\|\pi_2 u\| \leq 2de_2 \|u\|$. Hence,

$$\begin{aligned} \angle(T_0, T) &= \max_{u \in T} \max_{v' \in T_0^\perp} \frac{|\langle u, v' \rangle|}{\|u\| \|v'\|} \\ &\leq 2de_2. \end{aligned}$$

□

E Local PCA for Tangent Space Estimation

This section is dedicated to the proofs of Section 5. The models with and without outliers are considered jointly as often as possible. $\wedge$ and $\vee$ denote respectively the minimum and the maximum of real numbers. We first state elementary results which will be combined to prove our main results.

E.1 Preliminary results

We will restrict our attention to points X_j that are close enough to M so that $B(X_j, h) \cap M$ has enough probability mass. To this aim, we adopt the following notation. For a fixed point x, let $p(x, h)$ denote $P(\mathcal{B}(x, h))$. We decompose it as $p(x, h) = \beta q(x, h) + (1 - \beta) q'(x, h)$, where $q(x, h) = Q(\mathcal{B}(x, h))$, and $q'(x, h) = (h/K_0)^D$.

Lemma 33. *There exists $h_+(\rho, \beta, f_{min}, f_{max}, d)$ such that, if $h \leq h_+$, for every x such that $\mathrm{d}(x, M) \leq h$, we have*

- $\mathcal{B}(x, h) \cap M \subset \mathcal{B}(\pi_M(x), 2h) \cap M$,
- $q(x, h) \leq C_d f_{max} h^d$.

Moreover, if $\mathrm{d}(x, M) \leq h/\sqrt{2}$, we have

- $\mathcal{B}(\pi_M(x), h/(2\sqrt{2})) \cap M \subset \mathcal{B}(x, h)$,
- $c_d f_{min} h^d \leq q(x, h)$,
- $p(x, h) \leq 2\beta q(x, h)$.

At last, we may choose $h_+ \leq \rho/\sqrt{3d}$.

Proof of Lemma 33. Set $h_1(\rho) = \rho/(12\alpha)$, and let x be such that $\mathrm{d}(x, M) \leq h$, and $h \leq h_1$. According to Proposition 26 and Proposition 28, $\mathcal{B}(x,h) \cap M \subset \mathcal{B}(\pi_M(x), r_h^+) \cap M$, with $r_h^+ = (1 + 4\alpha^2\Delta/\rho)r_h \leq 2r_h$. According to Proposition 27, we also have $q(x,h) \leq C_d f_{max} h^d$.

Now if $\mathrm{d}(x, M) \leq h/\sqrt{2}$, $\mathcal{B}(\pi_M(x), r_h^-) \cap M \subset M$ according to Proposition 28, with $r_h^- = (1 - \alpha^2\Delta/\rho)r_h \geq r_h/2 \geq h/(2\sqrt{2})$. According to Proposition 27, we may write $c_d f_{min} h^d \leq q(x,h)$.

Applying Proposition 27 again , there exists $h_2(f_{min}, d, D, \beta, \rho)$ such that if $h \leq h_1 \wedge h_2$, then for any x such that $\mathrm{d}(x, M) \leq h/\sqrt{2}$ we have $(1-\beta)q'(x,h) \leq \beta C_{d,f_{min},\rho} h^d$, along with $p(x) \leq 2\beta q(x)$. Taking $h_+ = h_1 \wedge h_2 \wedge \rho/\sqrt{3d}$ leads to the result. □

Without loss of generality, the local PCA analysis will be conducted for $X_1 \in \mathbb{X}^{(k)}$, for some fixed $k \geq -1$, the results on the whole sample then follow from a standard union bound. For convenience, we assume that $\pi_M(X_1) = 0$ and that T_0M is spanned by the d first vectors of the canonical basis of $\mathbb{R}^D$. In what follows, denote by $\hat{t}$ the map from $\mathbb{R}^D$ to $\{0,1\}$ such that $\hat{t}(X_i) = 1$ if and only if X_i is in $\mathbb{X}^{(k)}$.

Let $U(X_i, h)$, $i = 2, \ldots, n$, denote $\mathbb{1}_{\mathcal{B}(X_1,h)}(X_i)$, and let $V_{2:n}$ denote the vector such that $V_i = 1$ if X_i is drawn from the noise distribution. It is immediate that the $(U_i(h), V_i)$'s are independent and identically distributed, with common law $(U(h), V)$.

With a slight abuse of notation, we will denote by $\mathbb{P}$ and $\mathbb{E}$ conditional probability and expectation with respect to X_1. The following expectations will be of particular interest.

$$\begin{aligned} m(h) &= \mathbb{E}(XU(X,h)V)/\mathbb{E}(VU(X,h)), \\ \Sigma(h) &= \mathbb{E}(X - m(h))_\top (X - m(h))_\top^t U(X,h) V, \end{aligned}$$

where for any x in $\mathbb{R}^D$ $x_\top$ and $x_\perp$ denote respectively the projection of x onto T_0M and $T_0M^\perp$.

The following Lemma gives useful results on both $m(h)$ and $\Sigma(h)$, provided that X_1 is close enough to M.

Lemma 34. *If $d(X_1, M) \leq h/\sqrt{2}$, for $h \leq h_+$, then*

$$\Sigma(h) = \left(\begin{array}{c|c} A(h) & 0 \\ \hline 0 & 0 \end{array}\right),$$

with

$$\lambda_{min}(A(h)) \geq \beta c_{d,f_{min},f_{max}} h^{d+2}.$$

Furthermore,

$$\begin{aligned} \|m_\top(h)\| &\leq 2h, \\ \|m_\perp(h)\| &\leq \frac{2h^2}{\rho}. \end{aligned}$$

Proof of Lemma 34. According to Proposition 25 combined with Proposition 26 and Proposition 27, we may write, for $h \leq h_+$ and y in $\mathcal{B}(X_1, h) \cap M$,

$$y = rv + R(r, v),$$

in local polar coordinates. According to Lemma 33, we have $\mathcal{B}(\pi_M(X_1), r_h^-) \cap M \subset \mathcal{B}(X_1, h) \cap M$. Let u be a unit vector in T_0M. Then $\langle u, rv + R(r,v) - m_\top(h)\rangle^2 \geq \langle u, rv - m_\top(h)\rangle^2 / 2 - 3R(r,v)^2$.

Hence we may write

$$\begin{aligned}
\langle Au, u\rangle &= \beta \int_{\mathcal{B}(X_1,h)\cap M} \langle u, rv + R(r,v) - m_\top(h)\rangle^2 J(r,v) f(r,v) dr dv \\
&\geq \beta f_{min} c_d \int_{r=0}^{r_h^-} \left[\langle u, rv - m_\top(h)\rangle^2 r^{d-1}/2 - 3r^4/(4\rho^2)\right] dr dv \\
&\geq \beta f_{min} c_d \left(\frac{(r_h^-)^{d+2}}{2d(d+2)} - \frac{3(r_h^-)^{d+4}}{4\rho^2(d+4)} \right) \\
&\geq \beta c_{d,f_{min}} h^{d+2},
\end{aligned}$$

according to Lemma 33. Since $\mathcal{B}(X_1,h)\cap M \subset \mathcal{B}(\pi_M(X_1),2h)\cap M$, Proposition 26 combined with Jensen's inequality gives the upper bounds on $\|m_\perp(h)\|$ and $\|m_\top(h)\|$. □

We adopt the following notation for the local covariance matrix based on $\mathbb{X}^{(k)}$.

$$\begin{array}{rcl}
\hat{\Sigma}^{(k)}(h) & = & \frac{1}{n-1}\sum_{j\geq 2}(X_j - \bar{X}^{(k)})(X_j - \bar{X}^{(k)})^t U(X_i,h)\hat{t}(X_i), \\
\bar{X}^{(k)} & = & \frac{1}{N^{(k)}}\sum_{i\geq 2} X_i U(X_i,h)\hat{t}(X_i) \\
N^{(k)} & = & \sum_{i\geq 2} U(X_i,h)\hat{t}(X_i).
\end{array}$$

Note that the tangent space estimator $\texttt{TSE}(\mathbb{X}^{(k)},h)_1$ is the space spanned by the first d eigenvectors of $\hat{\Sigma}^{(k)}$. At last we define $N_0(h)$ and $N_1(h)$ as the number of points drawn from respectively noise and signal in $\mathcal{B}(X_1,h)\cap M$, namely

$$\begin{aligned}
N_1(h) &= \sum_{i\geq 2} U(X_i,h)V_i, \\
N_0(h) &= \sum_{i\geq 2} U(X_i,h)(1-V_i).
\end{aligned}$$

To quantify the deviations of empirical means and covariances from their expectations, we need the following Lemma.

Lemma 35. *Recall that* $h_0 = \left(\kappa \frac{\log(n-1)}{\beta(n-1)}\right)^{1/(d+1)}$ *and* $h_\infty = h_0^{(d+1)/d}$, *for* κ *to be fixed later.*

If $h_0 \leq h_+$ *and* $d(X_1,M) \leq h_+/\sqrt{2}$, *then, with probability larger than* $1 - 6\left(\frac{1}{n-1}\right)^{2/d+1}$, *the following inequalities hold, for all* $h \leq h_0$ *and every possible slab* S_T.

$$\begin{array}{rcl}
\frac{N_0(h)}{n-1} & \leq & 2(1-\beta)q'(h) + \frac{10(2+2/d)\log(n-1)}{n-1}, \\
\frac{N_1(h)}{n-1} & \leq & 2\beta q(h) + \frac{10(2+2/d)\log(n-1)}{n-1}, \\
|S_T \cap \{X_2,\ldots,X_n\}|/(n-1) & \leq & 2PS_T + \frac{(10D_{VC}+5(1+1/d))\log(n-1)}{n-1}, \\
|S_T \cap \{X_2,\ldots,X_n\}|/(n-1) & \geq & PS_T - \frac{(2D_{VC}+1+2/d)\log(n-1)}{n-1},
\end{array}$$

where D_{VC} *is the Vapnik-Chervonenkis dimension of slabs in* $\mathbb{R}^D$, *and is therefore a fixed constant depending on* D.

Moreover, for all $(h_\infty \vee \sqrt{2}\mathrm{d}(X_1, M)) \leq h \leq h_0$*, and* n *large enough,*

$$\left\| \frac{1}{n-1} \sum_{i\geq 2} (X_i - m(h))_\top (X_i - m(h))_\top^t U(X_i, h) V_i - \Sigma(h) \right\|_{\mathcal{F}} \leq C_d \frac{f_{max}}{f_{min}\sqrt{\kappa}} \beta q(h) h^2,$$

$$\frac{1}{n-1} \left\| \sum_{i\geq 2} (X_i - m(h))_\top U(X_i, h) V_i \right\|_{\mathcal{F}} \leq C_d \frac{f_{max}}{f_{min}\sqrt{\kappa}} \beta q(h) h.$$

Proof of Lemma 35. The first two inequalities follows from Theorem 5.1 in [6]. The third and fourth inequalities proceed from the same result, as done in Lemma 7 of [23]. The two last results are derived from Talagrand-Bousquet's inequality (see, e.g., Theorem 2.3 in [8]) combined with the so-called peeling device.

Define $h_- = (h_\infty \vee \sqrt{2}\mathrm{d}(X_1, M))$, where we recall that in this analysis X_1 is fixed, and let $f_{T,h}$ denote the function

$$f_{T,h}(x, v) = \left\langle T, (x - m(h))_\top (x - m(h))_\top^t U(x, h) v \right\rangle,$$

for $h_- \leq h \leq h_0$, T a $d \times d$ matrix such that $\|T\|_{\mathcal{F}} = 1$, x in $\mathbb{R}^D$ and v in $\{0, 1\}$. Now we define the weighted empirical process

$$Z = \sup_{T,h} \quad \sum_{i\geq 2} \frac{f_{T,h}(X_i, V_i) - \mathbb{E} f_{T,h}(X, V)}{r(h)},$$

with $r(h) = \beta q(h) h^2$, along with the constrained empirical processes

$$Z(u) = \sup_{T, h\leq u} \quad \sum_{i\geq 2} f_{T,h}(X_i, V_i) - \mathbb{E} f_{T,h}(X, V),$$

for $h_- \leq u \leq h_0$. Since $\|f_{T,h}\|_\infty \leq 4h^2$ et $Var(f_{T,h}(X, V)) \leq 16\beta q(h) h^4$, a direct application of Theorem 2.3 in [8] yields, with probability larger than $1 - e^{-x}$,

$$Z(u) \leq 3\mathbb{E}Z(u) + \sqrt{\frac{32\beta q(u) u^4 x}{n-1}} + \frac{20u^2 x}{3(n-1)}.$$

To get a bound on $\mathbb{E}Z(u)$, we introduce some independent Rademacher random variables $\sigma_2, \ldots, \sigma_n$, i.e. $\mathbb{P}(\sigma_j = 1) = \mathbb{P}(\sigma_j = -1) = 1/2$, so that, according to the symmetrization principle (see, e.g.,

[27]), we have

$$
\begin{aligned}
(n-1)\mathbb{E}Z(u) &\le 2\mathbb{E}_{(X,V)_{2:n}}\mathbb{E}_\sigma \sup_{h\le u,T}\sum_{i\ge 2}\left\langle T, \sigma_i V_i U(X_i,h)((X_i-m(h))_\top (X_i-m(h))_\top^t)\right\rangle \\
&\le 2\mathbb{E}_{(X,V)_{2:n}}\mathbb{E}_\sigma \sup_{h\le u,T}\sum_{i\ge 2}\sigma_i\left\langle V_i U(X_i,h)X_i X_i^t, T\right\rangle \\
&\quad + 2\mathbb{E}_{(X,V)_{2:n}}\mathbb{E}_\sigma \sup_{h\le u,T}\sum_{i\ge 2}\sigma_i\left\langle V_i U(X_i,h)X_i m(h)^t, T\right\rangle \\
&\quad + 2\mathbb{E}_{(X,V)_{2:n}}\mathbb{E}_\sigma \sup_{h\le u,T}\sum_{i\ge 2}\sigma_i\left\langle V_i U(X_i,h)m(h)X_i^t, T\right\rangle \\
&\quad + 2\mathbb{E}_{(X,V)_{2:n}}\mathbb{E}_\sigma \sup_{h\le u,T}\sum_{i\ge 2}\sigma_i\left\langle V_i U(X_i,h)m(h)m(h)^t, T\right\rangle \\
&:= 2\mathbb{E}_{(X,V)_{2:n}}(E_1+E_2+E_3+E_4).
\end{aligned}
$$

where $\mathbb{E}_Y$ denotes expectations with respect to the random variable Y. For a fixed $(X,V)_{2:n}$, we may write

$$
\begin{aligned}
E_1 \le\ & \mathbb{E}_\sigma \sup_{h\le u}\left(\left\|\sum_{i\ge 2}\sigma_i V_i U(X_i,h)X_i X_i^t\right\|_{\mathcal{F}} - \mathbb{E}_\sigma\left\|\sum_{i\ge 2}\sigma_i V_i U(X_i,h)X_i X_i^t\right\|_{\mathcal{F}}\right) \\
&\qquad + \sup_{h\le u}\mathbb{E}_\sigma\left\|\sum_{i\ge 2}\sigma_i V_i U(X_i,h)X_i X_i^t\right\|_{\mathcal{F}} \\
:=\ & E_{11}+E_{12}.
\end{aligned}
$$

Jensen's inequality ensures that

$$
\begin{aligned}
E_{12} &\le \sup_{h\le u}\sqrt{\mathbb{E}_\sigma\left\|\sum_{i\ge 2}\sigma_i V_i U(X_i,h)X_i X_i^t\right\|_{\mathcal{F}}^2} \\
&\le 4u^2\sqrt{N_1(u)},
\end{aligned}
$$

hence

$$
\mathbb{E}_{X_{2:n}}E_{12} \le 4u^2\sqrt{\beta(n-1)q(u)}.
$$

For the remaining term E_{11}, note that, when $(X,V)_{2:n}$ is fixed,

$$
\sup_{h\le u}\left(\left\|\sum_{i\ge 2}\sigma_i V_i U(X_i,h)X_i X_i^t\right\|_{\mathcal{F}} - \mathbb{E}_\sigma\left\|\sum_{i\ge 2}\sigma_i V_i U(X_i,h)X_i X_i^t\right\|_{\mathcal{F}}\right)
$$

is in fact a supremum of at most $N_1(u)$ processes, each of them being subGaussian with variance bounded by $16h^4 N_1(u)$, according to a bounded difference inequality (see, e.g., Theorem 6.2 of [7] along with Section 2.3 of [7]). Hence a maximal inequality for subGaussian random variables (see Section 2.5 of [7]) ensures that

$$E_{11} \leq 4h^2\sqrt{2N_1(u)\log(N_1(u))} \leq 4h^2\sqrt{2N_1(u)\log(n-1)}.$$

Hence $\mathbb{E}_{(X,V)_{2:n}} E_{11} \leq 4h^2\sqrt{2\beta(n-1)q(u)\log(n-1)}$.

E_2 may also be decomposed as

$$\begin{aligned}
E_2 &= \mathbb{E}_\sigma \sup_{h\leq u} \left\| \left(\sum_{i\geq 2} \sigma_i V_i U(X_i,h) X_i \right) m(h)^t \right\|_{\mathcal{F}} \\
&\leq 2u\mathbb{E}_\sigma \sup_{h\leq u} \left\| \sum_{i\geq 2} \sigma_i V_i U(X_i,h) X_i \right\| \\
&\leq 2u \left(\mathbb{E}_\sigma \sup_{h\leq u} \left(\left\| \sum_{i\geq 2} \sigma_i V_i U(X_i,h) X_i \right\| - \mathbb{E}_\sigma \left\| \sum_{i\geq 2} \sigma_i V_i U(X_i,h) X_i \right\| \right) + \sup_{h\leq u} \mathbb{E}_\sigma \left\| \sum_{i\geq 2} \sigma_i V_i U(X_i,h) X_i \right\| \right) \\
&:= 2u(E_{21} + E_{22}).
\end{aligned}$$

Jensen's inequality yields that $E_{22} \leq 2u\sqrt{N_1(u)}$, and the same argument as for E_{11} (expectation of a supremum of $n-1$ subGaussian processes with variance bounded by $4u^2 N_1(u)$) gives $E_{22} \leq 2u\sqrt{2N_1(u)\log(n-1)}$. Hence

$$\mathbb{E}_{(X,V)_{2:n}} E_2 \leq 4u^2\sqrt{\beta(n-1)q(u)} \left(\sqrt{2\log(n-1)} + 1 \right).$$

Similarly, we may write

$$\mathbb{E}_{(X,V)_{2:n}} E_3 \leq 4u^2\sqrt{\beta(n-1)q(u)} \left(\sqrt{2\log(n-1)} + 1 \right).$$

At last, we may decompose E_4 as

$$\begin{aligned}
E_4 &\leq \mathbb{E}_\sigma 4u^2 \sup_{h\leq u} \left| \sum_{i\geq 2} V_i U(X_i,h) \right| \\
&\leq 4u^2 \left[\mathbb{E}_\sigma \sup_{h\leq u} \left(\left| \sum_{i\geq 2} V_i U(X_i,h) \right| - \mathbb{E}_\sigma \left| \sum_{i\geq 2} V_i U(X_i,h) \right| \right) + \sup_{h\leq u} \mathbb{E}_\sigma \left| \sum_{i\geq 2} V_i U(X_i,h) \right| \right] \\
&\leq 4u^2\sqrt{N_1(u)} \left(\sqrt{2\log(n-1)} + 1 \right),
\end{aligned}$$

using the same argument. Combining all these terms leads to

$$\mathbb{E}Z(u) \leq \frac{32\sqrt{\beta q(u)}}{\sqrt{n-1}} \left(\sqrt{2\log(n-1)} + 1 \right),$$

hence we get

$$\mathbb{P}\left(Z_u \geq \frac{192\sqrt{2}u^2\sqrt{\beta q(u)\log(n-1)}}{\sqrt{n-1}} \left(1 + \frac{1}{48}\sqrt{\frac{x}{\log(n-1)}} \right) + \frac{20u^2 x}{n-1} \right) \leq e^{-x}.$$

To derive a bound on the weighted process Z, we make use of the so-called peeling device (see, e.g., Section 5.3 of [31]). Set $p = \lceil \log(h_0/h_\infty) \rceil \leq 1 + \log(h_0/h_\infty)$, so that $e^{-p}h_0 \leq h_-$. According to Lemma 33, if I_j denotes the slice $[e^{-j}h_0, e^{-(j-1)}h_0] \cap [h_-, h_0]$, then, for every h in I_j, we have

$$r(h) \geq r(h_{j-1}) c_d \frac{f_{min}}{f_{max}},$$

where c_d depends only on the dimension, provided that $h_0 \leq h_+$. Now we may write

$$\begin{aligned}
&\mathbb{P}\left(Z \geq \frac{192 f_{max}\sqrt{2}}{f_{min} c_d \sqrt{\beta q(h_-)(n-1)}} \left(1 + \frac{1}{48}\sqrt{\frac{x+\log(p)}{n-1}}\right) + \frac{20 f_{max}(x+\log(p))}{(n-1)\beta c_d f_{min} q(h_-)} \right) \\
&\leq \sum_{j=1}^{p} \mathbb{P}\left(\sup_{T, h \in I_j} \frac{\sum_{i \geq 2} f_{T,h}(X_i, V_i) - \mathbb{E} f_{T,h}(X,V)}{r(h)} \geq \frac{192 f_{max}\sqrt{2}}{f_{min} c_d \sqrt{\beta q(h_-)(n-1)}} \left[1 + \frac{1}{48}\sqrt{\frac{x+\log(p)}{n-1}}\right] \right. \\
&\qquad \left. + \frac{20 f_{max}(x+\log(p))}{(n-1) f_{min} c_d \beta q(h_-)} \right) \\
&\leq \sum_{j=1}^{p} \mathbb{P}\left(Z(h_{j-1}) \geq \frac{192\sqrt{2} r(h_{j-1})}{\sqrt{\beta q(h_-)(n-1)}} \left(1 + \frac{1}{48}\sqrt{\frac{x+\log(p)}{n-1}}\right) + \frac{20 r(h_{j-1})(x+\log(p))}{(n-1)\beta q(h_-)} \right)
\end{aligned}$$

Since $q(h_{j-1}) \geq q(h_-)$, we deduce that

$$\begin{aligned}
\mathbb{P}\left(Z \geq \frac{192 f_{max}\sqrt{2}}{f_{min} c_d \sqrt{\beta q(h_-)(n-1)}} \left(1 + \frac{1}{48}\sqrt{\frac{x+\log(p)}{n-1}}\right) + \frac{20 f_{max}(x+\log(p))}{(n-1) c_d f_{min} \beta q(h_-)} \right) &\leq p e^{-(x+\log(p))} \\
&\leq e^{-x}.
\end{aligned}$$

Now, according to Lemma 33, $\beta q(h_-) \geq c_d \kappa \log(n-1)/(n-1)$. On the other hand, $p \leq 1 + \log(h_0/h_\infty) \leq \log(\beta(n-1)/\kappa)/d \leq \log(n-1)/d$, for $\kappa \geq 1$. Now, for n large enough, taking $x = (1+2/d)\log(n-1)$ in the previous inequality, we get

$$\mathbb{P}(Z \geq C_d \frac{f_{max}}{f_{min}\sqrt{\kappa}}) \leq \left(\frac{1}{n-1}\right)^{1+2/d}.$$

The last concentration inequality of Lemma 35 may be derived the same way, considering the functions

$$g_{T,h}(x,v)) = \langle (x - m(h)) U(x,h) v, T \rangle,$$

where T is an element of $\mathbb{R}^d$ satisfying $\|T\| \leq 1$. □

E.2 Rates of convergence for tangent space estimation

From now on we suppose that all the inequalities of Lemma 35 are satisfied, defining then a global event of probability larger than $1 - 6\left(\frac{1}{n-1}\right)^{1/d+1}$.

We recall that we consider $h_-(X_1) \leq h \leq h_k$, $k \geq -1$ (with $h_{-1} = h_0$), and X_1 in $\mathbb{X}^{(k)}$ such that $\mathrm{d}(X_1, M) \leq h/\sqrt{2}$. We may then decompose the local covariance matrix as follows.

$$\begin{aligned}
\hat{\Sigma}^{(k)} &= \frac{1}{n-1}\sum_{i\geq 2}(X_i - m(h))(X_i - m(h))^t U(X_i,h)\hat{t}(X_i) - \frac{N^{(k)}}{n-1}(\bar{X}^{(k)} - m(h))(\bar{X}^{(k)} - m(h))^t \\
&= \frac{1}{n-1}\sum_{i\geq 2}(X_i - m(h))(X_i - m(h))^t U(X_i,h)\hat{t}(X_i)V_i(X_i) \\
&\quad + \frac{1}{n-1}\sum_{i\geq 2}(X_i - m(h))(X_i - m(h))^t U_i(1-V_i)\hat{t}(X_i) \\
&\quad - \frac{N^{(k)}}{n-1}(\bar{X}^{(k)} - m(h))(\bar{X}^{(k)} - m(h))^t, \\
&:= \hat{\Sigma}_1^{(k)} + \hat{\Sigma}_2^{(k)} + \hat{\Sigma}_3^{(k)}. \qquad (2)
\end{aligned}$$

Now assume that $V_i = 1$ implies $\hat{t}(X_i) = 1$. This is true for $k = -1$, since in this case $\hat{t}$ is always equal to 1. For $k \geq 0$ this is part of the induction hypothesis. Then the first term may be written as

$$\begin{aligned}
\frac{1}{n-1}\sum_{i\geq 2}(X_i - m(h))(X_i - m(h))^t U(X_i,h)V_i &= \frac{1}{n-1}\sum_{i\geq 2}(X_i - m(h))_\top (X_i - m(h))_\top^t U(X_i,h)V_i + R_1 \\
&= \Sigma(h) + R_1 + R_2,
\end{aligned}$$

where

$$\Sigma(h) = \begin{pmatrix} A(h) & 0 \\ 0 & 0 \end{pmatrix}.$$

According to Lemma 34, $\lambda_{min}(A(h)) \geq c_d f_{min}\beta h^{d+2}$, and $\|R_1\|_{\mathcal{F}} \leq 3\frac{N_1(h)h^3}{\rho(n-1)}$ according to Proposition 26. Moreover, we can write

$$R_2 = \begin{pmatrix} R_2 & 0 \\ 0 & 0 \end{pmatrix},$$

with $\|R_2\|_{\mathcal{F}} \leq C_d \frac{f_{max}}{f_{min}\sqrt{\kappa}}\beta q(h)h^2$ according to Lemma 35.

Noise-Free Case In this section $\mathbb{X}^{(k)} = \mathbb{X}_n$ and $h_k = h_0$. Moreover, we can set $\beta = 1$, $\hat{t} = 1$, and $V_i = 1$ for all i in equation (2) to get $\hat{\Sigma}_2^{(k)} = 0$. The third term $\hat{\Sigma}_3^{(k)}$ may also be written as

$$\hat{\Sigma}_3^{(k)} = \begin{pmatrix} R_6 & 0 \\ 0 & 0 \end{pmatrix} + R_5,$$

with

$$\begin{aligned}
\|R_6\|_{\mathcal{F}} &\leq \frac{N^{(k)}}{n-1}\|(\bar{X}^{(k)} - m(h))_\top\|\|(\bar{X}^{(k)} - m(h))\| \\
&\leq \frac{2h}{n-1}\left\|\sum_{i\geq 2}(X_i - m)_\top U(X_i,h)\right\| \\
&\leq \frac{2C_d q(h)h^2 f_{max}}{f_{min}\sqrt{\kappa}},
\end{aligned}$$

according to Lemma 35. A similar bound on R_5 may be derived,

$$\begin{aligned}\|R_5\|_{\mathcal{F}} &\leq \frac{N^{(k)}}{n-1}\|(\bar{X}^{(k)} - m(h))_{\perp}\|(\bar{X}^{(k)} - m(h))\| \\ &\leq \frac{2h}{n-1}\left\|\sum_{i\geq 2}(X_i - m)_{\perp}U(X_i, h)\right\| \\ &\leq \frac{16N_1(h)h^3}{(n-1)\rho},\end{aligned}$$

according to Proposition 26. If we choose $h = \left(\kappa\frac{\log(n-1)}{n-1}\right)^{1/d}$, for κ large enough (depending on d, f_{min} and f_{max}), we have

$$\frac{\|R_2 + R_6\|_{\mathcal{F}}}{\lambda_{min}(A(h))} \leq 1/4.$$

Now, provided that $\kappa \geq 1$, according to Lemma 35, we may write

$$\frac{\|R_1 + R_5\|_{\mathcal{F}}}{\lambda_{min}(A(h))} \leq K_{f_{max},f_{min},d}h/\rho,$$

which, for n large enough, leads to

$$\angle(T_0M, \hat{T}_{X_1}M) \leq K_{f_{max},f_{min},d}h/\rho,$$

according to Theorem 32.

Clutter noise, Initialization step Now we set $k = -1$, $\mathbb{X}^{(k)} = \mathbb{X}_n$, $\hat{t} = 1$, $h = h_0$ and $\mathrm{d}(X_1, M) \leq h_0/\sqrt{2}$. Term $\hat{\Sigma}_2^{(k)}$ in inequality (2) may be bounded by

$$\|\hat{\Sigma}_2^{(k)}\|_{\mathcal{F}} \leq \frac{16h^2N_0(h)}{n-1}.$$

In turn, term $\hat{\Sigma}_3^{(k)}$ may be decomposed as

$$\frac{N^{(k)}}{n-1}(\bar{X}^{(k)} - m(h))(\bar{X}^{(k)} - m(h))^t = \begin{pmatrix} R_6 & 0 \\ 0 & 0\end{pmatrix} + R_5,$$

with

$$\begin{aligned}\|R_6\|_{\mathcal{F}} &\leq \frac{N^{(k)}}{n-1}\|(\bar{X}^{(k)} - m(h))_{\top}\|\|(\bar{X}^{(k)} - m(h))\| \\ &\leq \frac{2h}{n-1}\left(\|\sum_{i\geq 2}(X_i - m(h))_{\top}U(X_i, h)V_i\| + \|\sum_{i\geq 2}(X_i - m(h))_{\top}U(X_i, h)(1 - V_i)\|\right) \\ &\leq \frac{2C_d\beta q(h)h^2f_{max}}{f_{min}\sqrt{\kappa}} + \frac{4h^2N_0(h)}{n-1},\end{aligned}$$

according to Proposition 35. We may also write

$$\begin{aligned}\|R_5\|_{\mathcal{F}} &\leq \frac{N^{(k)}}{n-1}\|(\bar{X}^{(k)}-m(h))_{\perp}\|\|(\bar{X}^{(k)}-m(h))\| \\ &\leq \frac{2h}{n-1}\left(\|\sum_{i\geq 2}(X_i-m(h))_{\perp}U_h(X_i)V_i\|+\|\sum_{i\geq 2}(X_i-m(h))_{\perp}U_h(X_i)(1-V_i)\|\right) \\ &\leq \frac{N_1(h)h^3}{(n-1)\rho}+\frac{10N_0(h)h^2}{(n-1)},\end{aligned}$$

according to Proposition 26. As in the noise-free case, provided that κ is large enough (depending on d f_{min} and d_{max}), we have

$$\frac{\|R_2+R_6\|_{\mathcal{F}}}{\lambda_{min}(A(h))}\leq 1/4.$$

Since $(n-1)h_0^d=\frac{\kappa\log(n-1)}{\beta h}$, if we ask $\kappa\geq\rho$, then for n large enough we eventually get

$$\frac{\|\hat{\Sigma}_2^{(k)}+R_1+R_5\|_{\mathcal{F}}}{\lambda_{min}(A(h))}\leq K_{d,f_{min},f_{max},\beta}\frac{h_0}{\rho},$$

according to Proposition 35. Then, for n large enough such that $K_{d,f_{min},f_{max},\beta}\frac{h_0}{\rho}\leq 1/4$, Theorem 32 can be applied.

Clutter noise, Iteration step Now we assume that $k\geq 0$, and that $\mathrm{d}(X_i,M)\geq h_k^2/\rho$ implies $\hat{t}(X_i)=0$, with $h_k=\left(\kappa\frac{\log(n-1)}{\beta(n-1)}\right)^{\gamma_k}$, γ_k being between $1/(d+1)$ and $1/d$. Let $h_\infty\leq h\leq h_k$, and suppose that $\mathrm{d}(X_1,M)\leq h/\sqrt{2}$. We can decompose II as

$$\begin{aligned}&\frac{1}{n-1}\sum_{i\geq 2}(X_i-m(h))(X_i-m(h))^tU(X_i,h)(1-V_i)\hat{t}(X_i) \\ &\quad=\frac{1}{n-1}\sum_{i\geq 2}(X_i-m(h))_{\top}(X_i-m(h))_{\top}^tU(X_i,h)(1-V_i)\hat{t}(X_i)+R_3 \\ &\quad=\begin{pmatrix}R_4 & 0\\ 0 & 0\end{pmatrix}+R_3,\end{aligned}$$

with $\|R_4\|_{\mathcal{F}}\leq\frac{4N_0(h)h^2}{n-1}$ and $\|R_3\|\leq\frac{40N_0(h)hh_k^2}{(n-1)\rho}$, according to Proposition 29, for n large enough so that $h_0^2/\rho\leq h_\infty$. Term $\hat{\Sigma}_3^{(k)}$ may also be written as

$$\frac{N^{(k)}}{n-1}(\bar{X}^{(k)}-m(h))(\bar{X}^{(k)}-m(h))^t=\begin{pmatrix}R_6 & 0\\ 0 & 0\end{pmatrix}+R_5,$$

with

$$\begin{aligned}\|R_6\|_{\mathcal{F}} &\leq \frac{N^{(k)}}{n-1}\|(\bar{X}^{(k)} - m(h))_\top\|\|(\bar{X}^{(k)} - m(h))\| \\ &\leq \frac{2h}{n-1}\left(\|\sum_{i\geq 2}(X_i - m(h))_\top U(X_i,h)V_i\| + \|\sum_{i\geq 2}(X_i - m(h))_\top U(X_i,h)(1-V_i)\hat{t}(X_i)\|\right) \\ &\leq \frac{2C_d\beta q(h)h^2 f_{max}}{f_{min}\sqrt{\kappa}} + \frac{4h^2 N_0(h)}{(n-1)},\end{aligned}$$

according to Proposition 35 and Proposition 29. We may also write

$$\begin{aligned}\|R_5\|_{\mathcal{F}} &\leq \frac{N^{(k)}}{n-1}\|(\bar{X}^{(k)} - m(h))_\perp\|\|(\bar{X}^{(k)} - m(h))\| \\ &\leq \frac{2h}{n-1}\left(\|\sum_{i\geq 2}(X_i - m(h))_\perp U_h(X_i)V_i\| + \|\sum_{i\geq 2}(X_i - m(h))_\perp U_h(X_i)(1-V_i)\hat{t}(X_i)\|\right) \\ &\leq \frac{N_1(h)h^3}{(n-1)\rho} + \frac{20N_0(h)hh_k^2}{\rho(n-1)},\end{aligned}$$

according to Proposition 26 and 29. As done before, we may choose κ large enough (depending on d, f_{min} and f_{max}, but not on k) such that

$$\frac{\|R_2 + R_4 + R_6\|_{\mathcal{F}}}{\lambda_{min}(A(h))} \leq 1/4.$$

Now choose $h = h_{k+1} = \left(\kappa\frac{\log(n-1)}{\beta(n-1)}\right)^{(2\gamma_k+1)/(d+2)}$, with $\kappa \geq 1$, then we get, according to Proposition 35,

$$\begin{aligned}\frac{\|R_1 + R_3 + R_5\|_{\mathcal{F}}}{\lambda_{min}(A)(h_{k+1})} &\leq C_d\frac{f_{max}h_{k+1}}{\rho f_{min}} + \frac{C'_d}{\beta\rho f_{min}}\left(\kappa\frac{\log(n-1)}{\beta(n-1)}\right)^{\gamma_{k+1}+2\gamma_k-(2\gamma_k+1)+1} \\ &\leq K(f_{max}, f_{min}, d, \beta)\frac{h_{k+1}}{\rho},\end{aligned}$$

where $K(f_{max}, f_{min}, d, \beta)$ does not depend on k either. At last, since $K(f_{max}, f_{min}, d, \beta)\frac{h_{k+1}}{\rho} \leq K(f_{max}, f_{min}, d, \beta)\frac{h_0}{\rho} \leq 1/4$, for n large enough (not depending on k), we may apply Theorem 32.

Clutter Noise, random iteration $\hat{k}$

In this case, we have $\mathrm{d}(X_j, M) \leq h_\infty^2/\rho$, for every X_j in $\mathbb{X}^{(\hat{k})}$. The proof of Proposition 21 follows from the same calculation as above, replacing h_k^2/ρ by its upper bound h_∞^2/ρ and taking $h_{k+1} = h_\infty$.

E.3 Denoising rates

We recall that the slab $S(x, T, h)$ is the set of points y such that $\|\pi_T(y-x)\| \leq k_1 h$ and $\|\pi_{T^\perp}(y-x)\| \leq k_2 h^2$, k_1 and k_2 defined in Lemma 17, and where π_T denotes the orthogonal projection onto T. The proof of the denoising performance of one iteration is derived from the following Lemma.

Lemma 36. *For κ and n large enough, there exists a threshold t such that, for all $h_\infty \leq h \leq h_0$, and under the assumptions of Lemma 35, we have*

$$\begin{array}{rcl} X_1 \in M \quad \text{and} \quad \angle(T, T_{X_1}M) \leq Kh/\rho & \Rightarrow & |S(X_1,T,h) \cap \{X_2,\dots,X_n\}| \geq t\log(n-1), \\ d(X_1,M) \geq h^2/\rho \quad \text{and} \quad \angle\left(T, T_{\pi(X_1)}\right) M \leq Kh/\rho & \Rightarrow & |S(X_1,T,h) \cap \{X_2,\dots,X_n\}| < t\log(n-1) \\ d(X_1,M) \geq h/\sqrt{2} & \Rightarrow & |S(X_1,T,h) \cap \{X_2,\dots,X_n\}| < t\log(n-1) \end{array}$$

Proof. If $d(X_1,M) \geq h/\sqrt{2}$ or $d(X_1,M) \geq h^2/\rho$ and $\angle\left(T_{\pi(X_1)}M, T\right) \leq Kh/\rho$, then, according to Lemma 17,

$$PS(X_1,T,h) = \frac{(1-\beta)}{K_0^D}(k_1 h)^d k_2^{D-d} h^{2(D-d))} = \frac{(1-\beta)k_1^d k_2^{D-d}}{K_0^D} h^{2D-d}.$$

According to Lemma 35, since $h \leq h_0$, we may write

$$\begin{aligned} |S(X_1,T,h) \cap \{X_2,\dots,X_n\}|/(n-1) \leq & \frac{2(1-\beta)k_1^d k_2^{D-d}\kappa^{(d+2)/d}}{M^D}\left(\frac{\log(n-1)}{\beta(n-1)}\right)^{1+1/(d+1)} \\ & + \frac{(20D_{VC} + 10(1+1/d))\log(n-1)}{n-1}. \end{aligned}$$

On the other hand, if $X_1 \in M$, we have $PS(X_1,T,h) \geq q(k_3 h)$, according to Lemma 17. Using concentration bounds again yields that

$$|S(X_1,T,h) \cap \{X_2,\dots,X_n\}|/(n-1) \geq (c_d f_{min}\kappa - (4D_{VC} + 2(1+1/d))\frac{\log(n-1)}{n-1}.$$

The result follows easily. □

Now if we choose $K = K(f_{max}, f_{min}, d, \beta)$ and the sequence of bandwidths h_k mentioned in the last section to construct the slabs, Proposition 18 follows straightforwardly, with a union bound on the sample $\{X_1,\dots,X_n\}$.

F Proof of the Main Reconstruction Results

We now prove main results Theorem 6 in the noise-free model, and Theorems 7 and 8 in the clutter noise model.

F.1 Noise-Free Case

Proof of Corollary 15. Let $Q \in \mathcal{G}_{D,d,f_{min},f_{max},\rho}$. Write $\varepsilon = c_{d,f_{min},f_{max}}\left(\frac{\log n}{n-1}\right)^{1/d}$. Consider the event A defined by

$$A = \left\{ \max_{X_j \in \mathbb{Y}_n} \angle(T_{X_j}M, \hat{T}_j) \leq C_{d,f_{min},f_{max}} \frac{\varepsilon}{\rho} \right\} \cap \left\{ \mathrm{d_H}(M, \mathbb{Y}_n) \leq 2\varepsilon \right\} \cap \left\{ \mathbb{Y}_n \text{ is } \varepsilon\text{-sparse} \right\}.$$

The construction of $\mathbb{Y}_n$ from the farthest point sampling algorithm and a straightforward combination of Proposition 14 and Lemma 30 yields, for n large enough,

$$\mathbb{P}_Q\left(A^c\right) \leq 7\left(\frac{1}{n-1}\right)^{1/d}.$$

□

Proof of Theorem 6. Following the above notation, we observe that on the event A, Theorem 13 holds with parameters $\tau_1 = 0$ and $\tau_2 = C_{d,f_{min},f_{max}}/\rho$, so that the first part of Theorem 6 is proved. Furthermore, for n large enough,

$$\begin{aligned}\mathbb{E}_Q\left[\mathrm{d_H}\left(M, \hat{M}_{\texttt{TDC}}\right)\right] &\leq \mathbb{E}_Q\left[\mathrm{d_H}\left(M, \hat{M}_{\texttt{TDC}}\right)\mathbb{1}_A\right] + \mathbb{E}_Q\left[\mathrm{d_H}\left(M, \hat{M}_{\texttt{TDC}}\right)\mathbb{1}_{A^c}\right] \\ &\leq (C_{d,\rho'_{\rho,\tau_1,\tau_2}} + \tau_2)\varepsilon^2 + \mathrm{diam}(M)\mathbb{P}_Q(A^c) \\ &\leq C'_{d,f_{min},f_{max},\rho}\varepsilon^2,\end{aligned}$$

where for the last line we used the diameter bound of Proposition 2. □

F.2 Clutter Noise Case

Proof of Corollary 19. Let $P \in \mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}$. For n large enough, Denote by ε the quantity $c_{d,f_{min},f_{max},\rho}\left(\frac{\log(n-1)}{\beta(n-1)}\right)^{\gamma_k}$. Consider the event

$$\begin{aligned}A^\delta = \left\{\max_{X_j \in \mathbb{Y}_n^\delta} \angle(T_{\pi(X_j)}M, \hat{T}_j) \leq C_{d,f_{min},f_{max},\rho}\varepsilon\right\} \cap \left\{\max_{X_j \in \mathbb{Y}_n^\delta} \mathrm{d}(X_j, M) \leq \varepsilon^2\right\} \\ \cap \left\{\mathrm{d_H}(M, \mathbb{Y}_n^\delta) \leq 2\varepsilon\right\} \cap \left\{\mathbb{Y}_n^\delta \text{ is } \varepsilon\text{-sparse}\right\}.\end{aligned}$$

From Lemma 30, Proposition 18 and the construction of $\mathbb{Y}_n^+$ with the farthest point sampling algorithm, it is clear that for n large enough,

$$\mathbb{P}_P\left(\left(A^\delta\right)^c\right) \leq 6\left(\frac{1}{n-1}\right)^{2/d} + \left(\frac{1}{n}\right)^{2/d} \leq 7\left(\frac{1}{n-1}\right)^{2/d}.$$

□

Proof of Theorem 7. Following the above notation, we observe that on the event A^+, Theorem 13 holds with parameters $\tau_1 = 1$ and $\tau_2 = C_{d,f_{min},f_{max},\rho}$, so that the first part of Theorem 7 is proved. Furthermore, for n large enough,

$$\begin{aligned}\mathbb{E}_P\left[\mathrm{d_H}\left(M, \hat{M}_{\texttt{TDC}\delta}\right)\right] &\leq \mathbb{E}_P\left[\mathrm{d_H}\left(M, \hat{M}_{\texttt{TDC}\delta}\right)\mathbb{1}_{A^\delta}\right] + \mathbb{E}_P\left[\mathrm{d_H}\left(M, \hat{M}_{\texttt{TDC}\delta}\right)\mathbb{1}_{(A^\delta)^c}\right] \\ &\leq (C_{d,\rho'_{\rho,\tau_1,\tau_2}} + \tau_1 + \tau_2)\varepsilon^2 + 2K_0\mathbb{P}_P\left(\left(A^\delta\right)^c\right) \\ &\leq C'_{d,f_{min},f_{max},\rho,\beta}\varepsilon^2,\end{aligned}$$

where for the second line we used Theorem 13 and the fact that $M \cup \hat{M}_{\texttt{TDC}\delta} \subset \mathcal{B}_0$, a ball of radius K_0. □

Theorem 8 is obtained similarly, using Proposition 21.