Minimax Boundary Estimation and Estimation with Boundary

Eddie Aamari, Catherine Aaron, Clément Levrard

Minimax Boundary Estimation and Estimation with Boundary

Eddie Aamari* Catherine Aaron† Clément Levrard‡

Abstract

We derive non-asymptotic minimax bounds for the Hausdorff estimation of d-dimensional submanifolds $M \subset \mathbb{R}^D$ with (possibly) non-empty boundary ∂M. The model reunites and extends the most prevalent $\mathcal{C}^2$-type set estimation models: manifolds without boundary, and full-dimensional domains. We consider both the estimation of the manifold M itself and that of its boundary ∂M if non-empty. Given n samples, the minimax rates are of order $O\big((\log n/n)^{2/d}\big)$ if $\partial M = \emptyset$ and $O\big((\log n/n)^{2/(d+1)}\big)$ if $\partial M \neq \emptyset$, up to logarithmic factors. In the process, we develop a Voronoi-based procedure that allows to identify enough points $O\big((\log n/n)^{2/(d+1)}\big)$-close to ∂M for reconstructing it. Explicit constant derivations are given, showing that these rates do not depend on the ambient dimension $D \gg d$.

1 Introduction

Topological data analysis and geometric inference techniques have significantly grown in importance in the high-dimensional statistics area, both in its theoretical and practical aspects [48, 19]. Unlike Lasso-type methods [33] which strongly rely on a specific coordinate system, geometric inference techniques naturally yield features that are invariant through rigid transformations of the ambient space.

A central problem in this field is manifold estimation [8, 31, 30, 3, 43, 29]. Assuming that data $\mathbb{X}_n = \{X_1, \dots, X_n\}$ originate from some unknown distribution P on $\mathbb{R}^D$, these works study the estimation of its support $M = \mathrm{Supp}(P) \subset \mathbb{R}^D$, assumed to be a submanifold of dimension $d \ll D$. This provides a non-linear dimension reduction, that can allow to mitigate the curse of dimensionality, and helps for data visualization [37]. Manifold estimation is also of crucial importance for inferring other geometric features of M, as it appears as a critical intermediate step in a growing series of plugin strategies. See for instance [18] for persistent homology, [9] for the reach, or [24] for density estimation.

1.1 Support estimation

Overview

So far, the statistical study of support estimation in Hausdorff distance has been carried out within two somehow orthogonal settings: Full dimensional domains $\dim(M) = D$ on one hand — which necessarily have non-empty boundary $\partial M \neq \emptyset$ —, and low-dimensional submanifolds $\dim(M) = d < D$ without boundary $\partial M = \emptyset$ on the other hand. More precisely:

*CNRS & U. Paris Cité & Sorbonne U., Paris, France (https://perso.lpsm.paris/~aamari/)
†CNRS & U. Clermont Auvergne, Clermont-Ferrand, France (https://lmbp.uca.fr/~aaron/)
‡CNRS & U. Paris Cité & Sorbonne U., Paris, France (http://www.normalesup.org/~levrard/)

(i) Assuming that $M = \mathrm{Supp}(P) \subset \mathbb{R}^D$ is *full-dimensional* $\dim(M) = D$ (i.e. roughly everywhere of non-empty interior) and that P has enough mass in every neighborhood of its support, [26, 20] derive error bounds of order $(\log n/n)^{1/D}$. Here, a rate-optimal estimator simply consists of the sample set $\hat{M} = \mathbb{X}_n$ itself. Even under the additional geometric restriction of M being convex, this rate is still the best possible, due to the possible outward corners a convex set may contain. Beyond convexity, faster rates can actually be attained with additional smoothness constraints: if the (topological) boundary $\bar{\partial}M$ of the convex M is $\mathcal{C}^2$-smooth, [26] derives a convergence rate of order $(\log n/n)^{2/(D+1)}$ by considering the convex hull $\hat{M} = \mathrm{Hull}(\mathbb{X}_n)$, which also allows to estimate $\bar{\partial}M$ with $\bar{\partial}\hat{M}$ at the same rate. This phenomenon was also exhibited by [40, 45, 4] in similar convexity-type settings. Let us also mention the recent work of [17], which proposed a computationally efficient (yet unfortunately rate-suboptimal) boundary labelling method.

Note that despite a nearly quadratic gain in the rate for smooth cases, this framework still remains a hopeless scenario for high dimensional datasets, as it heavily suffers from the curse of dimensionality, both statistically and computationally.
This paper extends these results for M possibly of lower dimension $d \ll D$ and curved.

(ii) To overcome the curse of dimensionality, assuming that $M = \mathrm{Supp}(P) \subset \mathbb{R}^D$ is a $\mathcal{C}^2$ submanifold of dimension $\dim(M) = d < D$ *with empty (differential) boundary* $\partial M = \emptyset$, [31, 36] show that the minimax rate of estimation of M is of order $(\log n/n)^{2/d}$. The estimator of [31] being intractable in practice, [2] later proposed an optimal algorithm that outputs a triangulation of the data points which is computable in polynomial time. Using local polynomials, faster estimation rates of order $(\log n/n)^{k/d}$ were also shown to be achievable over $\mathcal{C}^k$-smooth submanifolds [3]. Although insensitive to the ambient dimension $D \gg d$, these results highly rely on the fact that $\partial M = \emptyset$.
This paper extends these results for M possibly with non-empty boundary.

Background

By definition, a submanifold $M \subset \mathbb{R}^D$ of dimension d is a smooth subspace that can be parametrized locally by $\mathbb{R}^d$. Hence, neighborhoods of points in M all look like d-dimensional balls. In contrast, a manifold *with boundary* is a smooth space that can be parametrized locally by $\mathbb{R}^d$ or by $\mathbb{R}^{d-1} \times \mathbb{R}_+$. If not empty, the boundary of M, denoted by ∂M, is the set of points nearby which M can only be parametrized by $\mathbb{R}^{d-1} \times \mathbb{R}_+$. Informally, the class of manifolds with boundary allows to take into account the possible "rims" a surface may contain (see Figure 1).

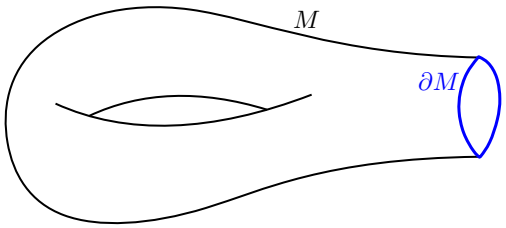


Figure 1: A surface M with non-empty boundary ∂M. Note that $\dim(\partial M) = \dim(M) - 1$, so that sample points from some roughly uniform distribution P on M almost surely never belong to ∂M. However, points close to ∂M should be processed differently in the analysis of such a sample, since they have an unbalanced neighborhood: they may cause boundary effects.

As mentioned above, most of the existing manifold estimation techniques require that $\partial M = \emptyset$,

which is very restrictive in view of real data [48]. When the empty boundary condition is dropped, the location of ∂M is often assumed to be known via an *oracle*, able to correctly label points that lie close to the boundary [44]. Prior to the present paper, a theoretically grounded construction of such an oracle given unlabeled data was not known, since the optimal detection and estimation rates of ∂M in arbitrary dimension had not been studied. This is mainly due to the technicalities that the presence of a boundary usually gives rise to. For instance, the restricted Delaunay triangulation to a surface with boundary may not even be homeomorphic to the surface [22]. Hence, Delaunay-based reconstructions are not good candidates to handle boundary, which contrasts sharply with the boundaryless case [11, 2]. Despite these barriers, a few interesting works on boundary inference can be found in the literature.

For surfaces in space ($d = 2$, $D = 3$), the so-called peeling algorithm consists in pruning an ambient triangulation (the α-shape of the point cloud) to handle boundary [22]. This method leverages boundary triangles being flatter than inner triangles. Unfortunately, such a method is limited to low dimensions, for the same instability problems described in [12].

On the other hand, in full dimension ($d = D$), [20] proposed a plugin estimator based on an estimator of M itself: under technical constraints, if $\hat{M}$ approximates M, then $\partial\hat{M}$ approximates ∂M. Such an plugin strategy provides a wide range of very general consistent boundary estimators: see for instance [45, 4] for convergence rates under additional assumptions. Note that naturally, such an approach is very costly — as acknowledged by the authors themselves —, and does not generalize easily to non-linear low-dimensional cases.

More recently, [5] designed an asymptotic boundary detection scheme based on local barycenter displacements: if a point $x \in M$ is close to ∂M, then the ball $M \cap \mathrm{B}(x, r)$ around x will not be balanced, and its barycenter would shift away from ∂M. This naturally yields a criterion to decide whether x belongs to ∂M or not. Unfortunately, this method requires the sampling density f over M to be Lipschitz and fails otherwise, as discontinuities of f far from ∂M may create artificial local barycenter shifts, and hence false positives. Let us also mention that this local barycenter shift has also been used in the context of density estimation on a manifold with boundary: [10] proposed a method for estimating the distance and direction of the boundary in order to correct the extra bias of a kernel density estimator near ∂M.

1.2 Contribution

This paper studies the minimax rates of estimation of d-dimensional $\mathcal{C}^2$-submanifolds $M \subset \mathbb{R}^D$ with possible $\mathcal{C}^2$ boundary ∂M (Definition 2.1), and the estimation of the boundary itself if not empty. As now standard in the literature, the loss is given by the Hausdorff distance $\mathrm{d_H}$ (a sup-norm between sets, see Definition 2.12), and $\mathcal{C}^2$ regularity of sets is measured through their reach $\tau_M, \tau_{\partial M} > 0$ (a generalized convexity parameter, see Definition 2.7).

Informally, we extend the known full-dimensional $\mathcal{C}^2$ support estimation rates to the case of low-dimensional curved M with $\mathcal{C}^2$ boundary. Indeed, if M is contained in a d-dimensional affine subspace of $\mathbb{R}^D$ and has a $\mathcal{C}^2$ boundary, its estimation boils down to the full dimensional case (Section 1.1 (i)), and can be done with rate $(\log n/n)^{2/(d+1)}$ [45, 4]. The present article proves that even if M is curved, the same rate drives the estimation hardness of M and ∂M. In addition, the estimator adapts automatically to the possible emptiness of ∂M, in which case M can be estimated at rate $(\log n/n)^{2/d}$ (see Section 1.1 (ii)). More precisely, we show that for n large

enough independent of D,

$$\inf_{\hat{B}_n} \sup_{\substack{\partial M \neq \emptyset \\ \tau_M \geq \tau_{\min} \\ \tau_{\partial M} \geq \tau_{\partial,\min}}} \vartheta_n(\partial M)^{-1} \mathbb{E}\left[\mathrm{d_H}\big(\partial M, \hat{B}_n\big)\right] = \tilde{\Theta}(1), \qquad \text{(Theorems 3.12 and 3.13)}$$

$$\inf_{\hat{M}_n} \sup_{\substack{\tau_M \geq \tau_{\min} \\ \tau_{\partial M} \geq \tau_{\partial,\min}}} \vartheta_n(\partial M)^{-1} \mathbb{E}\left[\mathrm{d_H}\big(M, \hat{M}_n\big)\right] = \tilde{\Theta}(1), \qquad \text{(Theorems 3.15 and 3.16)}$$

where $\hat{B}_n$ and $\hat{M}_n$ range among all the possible estimators based on n samples, and

$$\vartheta_n(\partial M) \asymp \begin{cases} (1/n)^{2/(d+1)} & \text{if } \partial M \neq \emptyset, \\ (1/n)^{2/d} & \text{if } \partial M = \emptyset. \end{cases}$$

These rates, given up to $\log n$ factors through $\tilde{\Theta}(1)$, do not depend on D.

1.3 Outline

We first describe the geometric framework and statistical setting we consider (Section 2). Then, we state the main boundary detection and estimation results (Section 3) and discuss them (Section 4). We present the principal steps of the proofs (Section 5). Finally, we discuss complexity, heuristics for parameter selection, and provide illustrations of the method on synthetic data (Section 6). For space constraints, the minor intermediate lemmas and most technical parts of the proofs are deferred to the Appendix.

2 Framework

Throughout, $D \geq 1$ is referred to as the ambient dimension and $\mathbb{R}^D$ is endowed with the Euclidean inner product $\langle \cdot, \cdot \rangle$ and the associated norm $\|\cdot\|$. The closed Euclidean ball of center x and radius r is denoted by $\mathrm{B}(x, r)$, and its open counterpart by $\mathring{\mathrm{B}}(x, r)$. Given a linear subspace $T \subset \mathbb{R}^D$, we also write $\mathrm{B}_T(0, r) := T \cap \mathrm{B}(0, r)$ for the r-ball of T centered at $0 \in T$.

2.1 Geometric setting

2.1.1 Submanifolds with boundary

By definition, the d-dimensional submanifolds $M \subset \mathbb{R}^D$ with boundary are the subsets of $\mathbb{R}^D$ that can locally be parametrized either by the Euclidean space $\mathbb{R}^d$, or the half-space $\mathbb{R}^{d-1} \times \mathbb{R}_+$ [38, Chapter 2].

Definition 2.1 (Submanifold with Boundary, Boundary, Interior)**.** A closed subset $M \subset \mathbb{R}^D$ is a d-dimensional $\mathcal{C}^2$-*submanifold with boundary* of $\mathbb{R}^D$, if for all $p \in M$ and all small enough open neighborhood V_p of p in $\mathbb{R}^D$, there exists an open neighborhood U_0 of 0 in $\mathbb{R}^D$ and a $\mathcal{C}^2$-diffeomorphism $\Psi_p : U_0 \to V_p$ with $\Psi_p(0) = p$, such that either:

(i) $\Psi_p\left(U_0 \cap \left(\mathbb{R}^d \times \{0\}^{D-d}\right)\right) = M \cap V_p$.
Such a $p \in M$ is called an *interior* point of M, the set of which is denoted by $\operatorname{Int} M$.

(ii) $\Psi_p\left(U_0 \cap \left(\mathbb{R}^{d-1} \times \mathbb{R}_+ \times \{0\}^{D-d}\right)\right) = M \cap V_p$.
Such a $p \in M$ is called a *boundary* point of M, the set of which is denoted by ∂M.

Remark 2.2 (Boundaries)**.** The *geometric* (or *differential*) boundary ∂M is not to be confused with the ambient *topological* boundary defined as $\bar{\partial} S := \bar{S} \setminus \mathring{S}$ for $S \subset \mathbb{R}^D$, where the closure and interior are taken with respect to the ambient topology of $\mathbb{R}^D$. Indeed, one easily checks that if $d < D$, then $\bar{\partial} M = M$. On the other hand, the two sets $\bar{\partial} M$ and ∂M coincide when $d = D$.

Then, submanifolds *without* boundary are those M that fulfill $\partial M = \emptyset$, i.e. that are everywhere locally parametrized by $\mathbb{R}^d$, and nowhere by $\mathbb{R}^{d-1} \times \mathbb{R}_+$. From this perspective — as confusing as this standard terminology can be —, submanifolds without boundary are special cases of submanifolds with boundary. Note that key instances of manifolds without boundary are given by boundaries of manifolds, as expressed by the following result.

Proposition 2.3 ([35, p.30])**.** *If $M \subset \mathbb{R}^D$ is a d-dimensional $\mathcal{C}^2$-submanifold with nonempty boundary ∂M, then ∂M is a $(d-1)$-dimensional $\mathcal{C}^2$-submanifold without boundary.*

Remark 2.4. If non-empty, this fact will allow us to estimate ∂M using the estimator designed for manifolds without boundary from [2], that we will build on top of some preliminarily filtered *boundary observations* (see Section 3.1).

2.1.2 Tangent and normal structures

In the present $\mathcal{C}^2$-smoothness framework, the difference between boundary and interior points sharply translates in terms of local first order approximation properties of M either by its so-called tangent cones or tangent spaces, which we now define (see Figure 2).

Definition 2.5 (Tangent and Normal Cones and Spaces)**.** Let $p \in M$, and Ψ_p its local parametrization from Definition 2.1.

- The *tangent cone* $Tan(p, M)$ of M at p is defined as

$$Tan(p, M) := \begin{cases} d_0\Psi_p(\mathbb{R}^d \times \{0\}^{D-d}) & \text{if } p \in \operatorname{Int} M, \\ d_0\Psi_p\left(\mathbb{R}^{d-1} \times \mathbb{R}_+ \times \{0\}^{D-d}\right) & \text{if } p \in \partial M, \end{cases}$$

  where $d_0\Psi_p$ denotes the differential of Ψ_p at 0.
  The *tangent space* T_pM is then defined as the linear span $T_pM := \operatorname{span}(Tan(p, M))$.

- The *normal cone* $Nor(p, M)$ of M at p is the dual cone of $Tan(p, M)$:

$$Nor(p, M) := \{v \in \mathbb{R}^D \mid \forall u \in Tan(p, M), \langle u, v \rangle \leq 0\}.$$

  The *normal space* of M at p is defined accordingly by $N_p(M) := \operatorname{span}(Nor(p, M))$.

Whenever $p \in \operatorname{Int} M$, it falls under the intuition that $Tan(p, M) = T_pM$ and $Nor(p, M) = N_pM$, while when $p \in \partial M$, N_pM and T_pM share one direction which is orthogonal to $T_p\partial M$. These properties are summarized in the following proposition (see Figure 2).

Proposition 2.6 (Outward-Pointing Vector)**.** *Let M be a $\mathcal{C}^2$-submanifold with boundary.*

- *If $p \in \operatorname{Int} M$, then $Tan(p, M) = T_pM$ and $Nor(p, M) = N_pM$ are orthogonal linear spaces spanning $\mathbb{R}^D$.*

- *If $p \in \partial M$, then $Tan(p, M)$ and $Nor(p, M)$ are complementary half-spaces, in the sense that $T_pM + N_pM = \mathbb{R}^D$ and $T_pM \cap N_pM$ is one-dimensional. The unique unit vector η_p in $Nor(p, M) \cap T_pM$ is called the* outward-pointing *vector. It satisfies*

$$Tan(p, M) = T_pM \cap \{\langle \eta_p, . \rangle \leq 0\}, \quad Nor(p, M) = N_pM \cap \{\langle \eta_p, . \rangle \geq 0\},$$

and

$$T_p\partial M \overset{\perp}{\oplus} \operatorname{span}(\eta_p) = T_pM,$$

where $\overset{\perp}{\oplus}$ denotes the orthogonal direct sum relation.

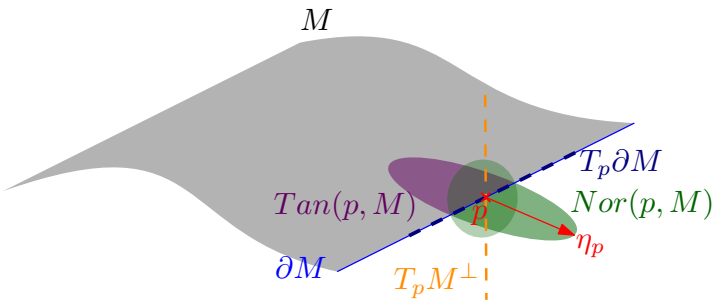


Figure 2: Tangent and normal structure of a surface ($d = 2$) in space ($D = 3$) at a boundary point.

The proof of Proposition 2.6 derives from elementary differential calculus and is omitted. The above purely differential definition of the tangent and normal cones coincides with that of the general framework of sets with positive reach [27] (to follow in Section 2.2). This general framework will enable us to quantify how well M is locally approximated by its tangent cones.

2.2 Geometric assumptions and statistical model

Any $\mathcal{C}^2$-submanifold M of $\mathbb{R}^D$ admits a tubular neighborhood in which any point has a unique nearest neighbor on M [15, p.93]. However, the width of this tubular neighborhood might be arbitrarily small. This scenario occurs when M exhibits high curvature or nearly self-intersecting areas [1]. In this case, the estimation of M gets more difficult, since such locations require denser sample to be reconstructed accurately. The width of such a tubular neighborhood is given by the so-called *reach* ([27, Defintion 4.1]), whose formal definition goes as follows.

Given a closed set $S \subset \mathbb{R}^D$, the *medial axis* $\operatorname{Med}(S)$ of S is the set of ambient points that do not have a unique nearest neighbor on S. More precisely, if

$$\mathrm{d}(z, S) := \min_{x \in S} \|z - x\|$$

stands for the *distance function* to S, then

$$\operatorname{Med}(S) := \left\{ z \in \mathbb{R}^D | \exists x \neq y \in S, \|z - x\| = \|z - y\| = \mathrm{d}(z, S) \right\}. \tag{1}$$

The reach of S is then defined as the minimal distance from S to $\operatorname{Med}(S)$.

Definition 2.7 (Reach)**.** The *reach* of a closed set $S \subset \mathbb{R}^D$ is

$$\tau_S := \min_{x \in S} \mathrm{d}\left(x, \operatorname{Med}(S)\right) = \inf_{z \in \operatorname{Med}(S)} \mathrm{d}\left(z, S\right).$$

By construction of the medial axis Equation (1), the projection on S

$$\pi_S(z) := \operatorname*{argmin}_{x \in S} \|x - z\|$$

is well defined (exactly) on $\mathbb{R}^D \setminus \mathrm{Med}(S)$. In particular, π_S is well defined on any r-neighborhood of S of radius $r < \tau_S$.

Remark 2.8. One easily checks that S is convex if and only if $\tau_S = \infty$ [27, Remark 4.2]. In particular, for the empty set $S = \emptyset$, we have $\tau_\emptyset = \infty$.

Requiring a lower bound on the reach of a manifold amounts to bound its curvature [42, Proposition 6.1], and prevents quasi self-intersection at scales smaller than the reach [1, Theorem 3.4]. Moreover, it allows to assess the quality of the linear approximation of the manifold by its tangent cones. In fact, [27, Theorem 4.18] shows that for all closed set $S \subset \mathbb{R}^D$ with reach $\tau_S > 0$, its tangent cone $Tan(x, S)$ is well defined at all $x \in S$, and $\mathrm{d}(y - x, Tan(x, S)) \leq \|y - x\|^2 / (2\tau_S)$ for all $y \in S$. This motivates the introduction of our geometric model below.

Definition 2.9 (Geometric Model)**.** Given integers $1 \leq d \leq D$ and positive numbers $\tau_{\min}, \tau_{\partial,\min}$, we let $\mathcal{M}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}$ denote the set of compact connected d-dimensional $\mathcal{C}^2$-submanifolds $M \subset \mathbb{R}^D$ with boundary, such that

$$\tau_M \geq \tau_{\min} \quad \text{and} \quad \tau_{\partial M} \geq \tau_{\partial,\min}.$$

Remark 2.10. Let us emphasize the following properties of the model:

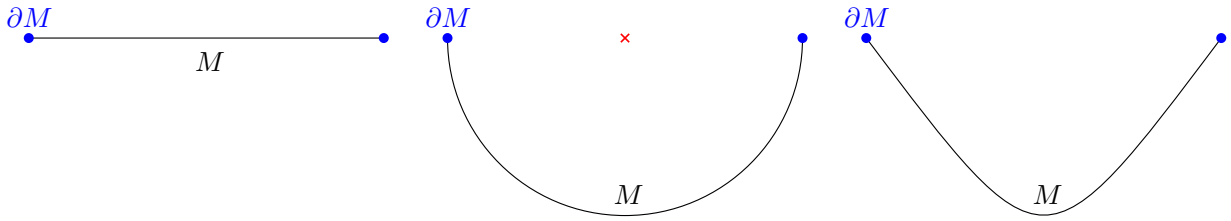
- The model $\mathcal{M}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}$ includes both submanifolds with empty and non-empty boundary ∂M, the main requirement being that $\tau_{\partial M} \geq \tau_{\partial,\min}$. If $\partial M = \emptyset$, this requirement is always fulfilled since $\tau_\emptyset = \infty$. Note also that Definition 2.9 does not exclude the case $d = D$, in which case M consists of a domain of $\mathbb{R}^D$ with non-empty interior. Furthermore, since the boundary ∂M of a submanifold M is either empty or itself a submanifold without boundary, a non-empty ∂M cannot be convex [34, Theorem 3.26]. As a result, $\mathcal{M}^{d,D}_{\tau_{\min},\infty}$ is exactly the set of submanifolds $M \in \mathcal{M}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}$ that have empty boundary. In particular, Definition 2.9 encompasses the model of [31, 36, 2].

- Similarly, since $\tau_M = \infty$ if and only if M is convex, $\mathcal{M}^{d,D}_{\infty,\tau_{\partial,\min}}$ is exactly the set of submanifolds $M \in \mathcal{M}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}$ that are convex (and hence have non-empty boundary). In particular, Definition 2.9 encompasses the model of [26].

- In full generality, the two lower bounds on the respective reaches of M and ∂M are *not* redundant with one another. As shown in Figure 3, τ_M and $\tau_{\partial M}$ are not related when $d < D$. However, for $d = D$, ∂M is the topological boundary of M (Remark 2.2). In this case, [27, Remark 4.2] and an elementary connectedness argument show that $\tau_M \geq \tau_{\partial M}$. Said otherwise, this means that the reach regularity of a full-dimensional domain is no worse than that of its boundary. Hence, $\mathcal{M}^{D,D}_{\tau_{\min},\tau_{\partial,\min}} = \mathcal{M}^{D,D}_{\tau_{\partial,\min},\tau_{\partial,\min}}$ for all $\tau_{\min} \leq \tau_{\partial,\min}$, so that for $d = D$, one may set $\tau_{\min} = \tau_{\partial,\min}$ without loss of generality.

The geometric model $\mathcal{M}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}$ being settled, we are now in position to define a generative model on such manifolds. In what follows, we let $\mathcal{H}^d$ denote the d-dimensional Hausdorff measure on $\mathbb{R}^D$ (see e.g. [28, Section 2.10.2]).

Definition 2.11 (Statistical Model)**.** Given $0 < f_{\min} \leq f_{\max} < \infty$, we let $\mathcal{P}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}(f_{\min}, f_{\max})$ denote the set of Borel probability distributions P on $\mathbb{R}^D$ such that:



(a) $\tau_{\partial M} < \tau_M = \infty$. (b) $\tau_{\partial M} = \tau_M$. (c) $\tau_{\partial M} > \tau_M$.

Figure 3: For $d < D$, the reach of a submanifold M and that of its boundary ∂M are not related.

- $M = \mathrm{Supp}(P) \in \mathcal{M}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}$,
- P has a density f with respect to the volume measure $\mathrm{vol}_M = \mathbb{1}_M \mathcal{H}^d$ on M, such that $f_{\min} \leq f(x) \leq f_{\max}$ for all $x \in M$.

From now on, we assume that we observe an i.i.d. n-sample $X_1, \ldots, X_n$ with unknown common distribution $P \in \mathcal{P}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}(f_{\min}, f_{\max})$, and denote the sample point cloud by

$$\mathbb{X}_n := \{X_1, \ldots, X_n\}.$$

Based on $\mathbb{X}_n$, the performance of the estimators of M and ∂M will be assessed in Hausdorff distance, which plays the role of a L^∞-distance between compact subsets of $\mathbb{R}^D$.

Definition 2.12 (Hausdorff Distance)**.** Given two compact subsets $S, S' \subset \mathbb{R}^D$, the *Hausdorff distance* between them is

$$\mathrm{d_H}(S, S') := \max\big\{\max_{x \in S} \mathrm{d}(x, S'), \max_{x' \in S'} \mathrm{d}(x', S)\big\}.$$

3 Main results

This section gathers the main results of this article: construction of estimators of ∂M and M, bounds on their Hausdorff performance, and nearly matching minimax lower bounds. To cope with the possible presence of a boundary, our first step is to determine which data points lie close to the boundary, if any.

3.1 Detecting boundary observations

3.1.1 Intuition

In the full-dimensional case ($d = D$), data points close to the boundary may be identified by how (macroscopically) large their Voronoi cells tend to be [45]. That is, if $\rho > 0$ is a detection radius, the *boundary observations* may be defined as

$$\mathcal{Y}_\rho = \{X_i \in \mathbb{X}_n \mid \exists O \in \mathbb{R}^D, \|O - X_i\| \geq \rho \text{ and } \mathring{\mathrm{B}}(O, \|O - X_i\|) \cap \mathbb{X}_n = \emptyset\}.$$

If X_i belongs to $\mathcal{Y}_\rho$ with associated $O \in \mathbb{R}^D$, then $\hat{\eta}_i := \frac{O - X_i}{\|O - X_i\|}$ appears to provide a consistent estimator of the unit outer normal vector of ∂M at $\pi_{\partial M}(X_i)$ [6]. The present work leverages the above intuition and extends it to the case where M is a d-dimensional manifold with $d < D$. In fact, the manifold M not being full-dimensional raises the following additional subtleties:

- Even if X_i is far from ∂M, its Voronoi cell is large in the directions of $T_{X_i}M^{\perp}$, as it actually contains at least $X_i + \mathrm{B}_{T_{X_i}M^{\perp}}(0, \tau_{\min})$. To detect points close to the boundary *only*, we shall hence avoid these normal non-informative directions and solely focus on the tangential components of the Voronoi cells. For instance, by first projecting points onto (an estimate of) $T_{X_i}M$.
- If X_i is close to ∂M but M is folded over X_i, then the Voronoi cell of X_i in the Voronoi diagram of the projected sample might be small (see Figure 4). To detect enough points close to the boundary, not all the sample should thus be projected, but rather just a neighborhood $\mathbb{X}_n \cap \mathrm{B}(X_i, R_0)$ of X_i, for some localization radius $R_0 > 0$ to be tuned.

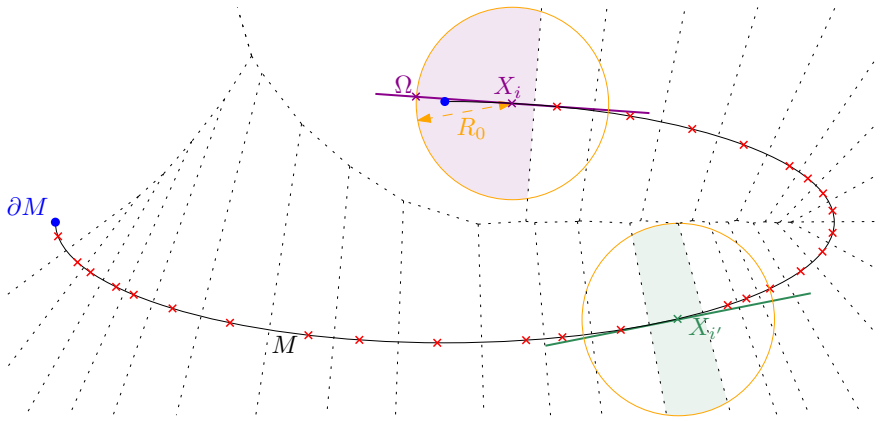


Figure 4: An ambient Voronoi diagram built on top of observations $\mathbb{X}_n$ lying on an open plane curve ($d = 1$, $D = 2$). The denser $\mathbb{X}_n$ in M, the narrower the Voronoi cell of the X_i's in the tangent directions $T_{X_i}M$. Observations close to ∂M yield cells that extend in the outward pointing direction. Localization radius $R_0 > 0$ prevents global foldings of M that would mix different ambient neighborhoods of M when projecting onto $T_{X_i}M$.

These two remarks lead to the following first detection procedure: for a collection of estimated tangent spaces $\hat{T}_i$'s, one *may* label X_i as being a *boundary observation* if it has a large Voronoi cell within its R_0-neighborhood, when projected onto $X_i + \hat{T}_i$. That is, if there exists $O \in \hat{T}_i$ such that $\|O\| \geq \rho$ and $\mathring{\mathrm{B}}(O, \|O\|) \cap \pi_{\hat{T}_i}(\mathbb{X}_n \cap \mathrm{B}(X_i, R_0) - X_i) = \emptyset$. Unfortunately, when $1 < d < D$, this intuitive detection method is not sufficient to provably detect enough observations close to the boundary. This issue can be overcome by investigating *all* the Voronoi cells of $\pi_{\hat{T}_j}(X_i)$ for $X_j \in \mathrm{B}(X_i, r) \cap \mathbb{X}_n$, where r is a small scale parameter. The details of this detection procedure are given in Section 3.1.3.

As it is now clear how critical the knowledge of tangent spaces is to build a Voronoi-based boundary detection scheme, let us first briefly detail how we estimate them.

3.1.2 Tangent space estimation

Following the ideas of [2], we will estimate tangent spaces using local principal component analysis.

Definition 3.1 (Tangent Space Estimator)**.** For $i \in \{1, \ldots, n\}$ and $h > 0$, we introduce the local covariance matrix

$$\hat{\Sigma}_i(h) := \frac{1}{n-1} \sum_{j \neq i} (X_j - X_i)(X_j - X_i)^t \mathbb{1}_{\mathrm{B}(X_i, h)}(X_j),$$

and define $\hat{T}_i$ as the linear span of the first d eigenvectors of $\hat{\Sigma}_i(h)$.

Note that $\hat{T}_i$ is a local estimator, in the sense that it is $\big((X_j - X_i)\mathbb{1}_{X_j \in \mathrm{B}(X_i,h)}\big)_{1\le j\le n}$-measurable (i.e. it only depends on the observations that are h-close to X_i). For a suitable choice of h, the following proposition provides guarantees on the principal angle between $T_{X_i}M$ and $\hat{T}_i$. In what follows, given two linear subspaces $T, T' \subset \mathbb{R}^D$, the *principal angle* between them is

$$\angle(T, T') := \|\pi_T - \pi_{T'}\|_{\mathrm{op}},$$

where $\|A\|_{\mathrm{op}} := \sup_{\|x\|\le 1} \|Ax\|$ stands for the operator norm of $A \in \mathbb{R}^{n\times n}$.

Proposition 3.2 (Tangent Space Estimation)**.** *Let $h = \big(C_d \frac{f_{\max}^4}{f_{\min}^5}\frac{\log n}{n-1}\big)^{\frac{1}{d}}$, for a large enough constant C_d. For n large enough so that $h \le \frac{\tau_{\min}}{32} \wedge \frac{\tau_{\partial,\min}}{3} \wedge \frac{\tau_{\min}}{\sqrt{d}}$, with probability larger than $1 - 2\big(\frac{1}{n}\big)^{\frac{2}{d}}$, we have*

$$\max_{1\le i\le n} \angle(T_{X_i}M, \hat{T}_i) \le C_d \frac{f_{\max}}{f_{\min}} \frac{h}{\tau_{\min}}.$$

A proof of Proposition 3.2 can be found in Appendix D.1. In what follows, we shall always choose h and n large enough as in Proposition 3.2.

3.1.3 Detection method and normal vector estimation

Now, for a local (though macroscopic) scale $R_0 > 0$, a detection radius $\rho > 0$ and a local bandwidth $r > 0$, we compute the d-dimensional Voronoi diagrams of $(\pi_{\hat{T}_i}(\mathrm{B}(X_i, R_0) \cap \mathbb{X}_n - X_i))_{1\le i\le n}$ and define our boundary observations detection procedure as follows. See Section 3.1.1 for a heuristic, and Figure 5 for an illustration associated with this definition.

Definition 3.3 (Boundary Observations)**.** For $i \in \{1, \dots, n\}$, we let $J_{R_0,r,\rho}(X_i)$ be the set of r-neighbors X_j of X_i for which X_i has a ρ-large Voronoi cell in the projected Voronoi diagram at X_j. That is, writing

$$\mathrm{Vor}_{R_0}^{(j)}(X_i) := \Big\{ O \in \hat{T}_j \,\Big|\, \mathring{\mathrm{B}}\big(O, \|O - \pi_{\hat{T}_j}(X_i - X_j)\|\big) \cap \pi_{\hat{T}_j}(\mathrm{B}(X_j, R_0) \cap \mathbb{X}_n - X_j) = \emptyset \Big\},$$

we define

$$J_{R_0,r,\rho}(X_i) := \Big\{ X_j \in \mathrm{B}(X_i, r) \cap \mathbb{X}_n \,\Big|\, \mathrm{Vor}_{R_0}^{(j)}(X_i) \cap \mathring{\mathrm{B}}_{\hat{T}_j}(\pi_{\hat{T}_j}(X_i - X_j), \rho)^c \neq \emptyset \Big\}.$$

The set of *boundary observations* $\mathcal{Y}_{R_0,r,\rho} \subset \mathbb{X}_n$ is then defined as the set of data points that have at least one such large Voronoi cell:

$$\mathcal{Y}_{R_0,r,\rho} := \{X_i \in \mathbb{X}_n \mid J_{R_0,r,\rho}(X_i) \neq \emptyset\}. \tag{2}$$

Remark 3.4. Detecting boundary observations requires to compute n Voronoi diagrams in dimension d. Note that this step does not depend on the ambient dimension D, and can run in parallel.

This strategy also provides a natural way to estimate unit normal outward-pointing vectors. For this, given a boundary observation $X_i \in \mathcal{Y}_{R_0,r,\rho}$, we simply consider directions in which $\mathrm{Vor}_{R_0}^{(j)}(X_i)$ is ρ-wide (see Figure 4). A formal definition goes as follows.

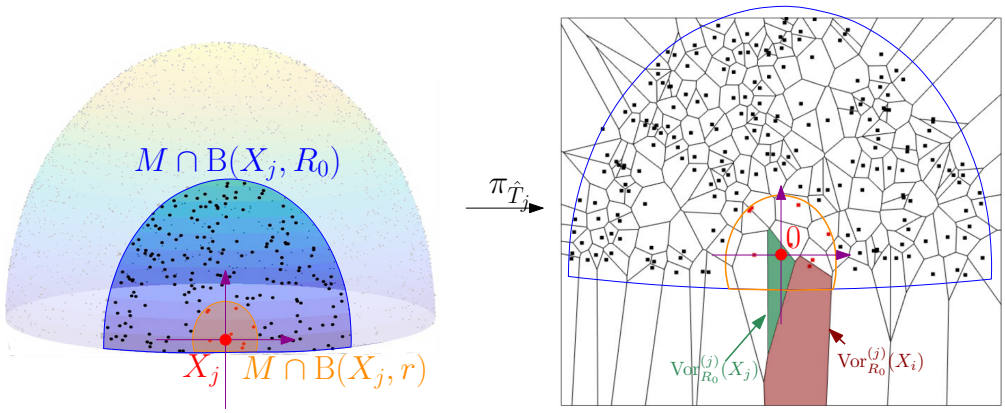


Figure 5: Illustration of Definition 3.3 over a half-sphere ($d = 2, D = 3$). Although the central point X_j (red) does not have a large Voronoi cell in $\hat{T}_j$, its neighbor X_i does. Therefore, X_j belongs to $J_{R_0,r,\rho}(X_i)$. In particular, X_i is labelled as a boundary point. Note that throughout the process, the Voronoi diagram is only computed in the d-planes $\{\hat{T}_j\}_{1\le j\le n}$, not in the ambient space $\mathbb{R}^D$.

Definition 3.5 (Normal Vector Estimator)**.** For $X_i \in \mathcal{Y}_{R_0,r,\rho}$ and $X_j \in J_{R_0,r,\rho}(X_i)$, let

$$\Omega^{(j)}_{R_0,r,\rho} \in \operatorname{argmin}\left\{ \left\| \Omega - \pi_{\hat{T}_j}(X_i - X_j) \right\| \,\middle|\, \Omega \in \mathrm{Vor}^{(j)}_{R_0}(X_i) \cap \mathring{\mathrm{B}}_{\hat{T}_j}(\pi_{\hat{T}_j}(X_i - X_j), \rho)^c \right\}.$$

The estimator of the unit normal outward-pointing vector in $\hat{T}_j$ is defined as

$$\tilde{\eta}^{(j)}_i := \frac{\Omega^{(j)}_{R_0,r,\rho} - \pi_{\hat{T}_j}(X_i - X_j)}{\left\| \Omega^{(j)}_{R_0,r,\rho} - \pi_{\hat{T}_j}(X_i - X_j) \right\|}.$$

The final estimator of the unit outward-pointing normal vector at X_i is then defined as

$$\tilde{\eta}_i := \frac{1}{\#J_{R_0,r,\rho}(X_i)} \sum_{j \in J_{R_0,r,\rho}(X_i)} \tilde{\eta}^{(j)}_i. \tag{3}$$

Remark 3.6. Let us mention that the choice of $\Omega^{(j)}_{R_0,r,\rho}$ in Definition 3.5 has been made to ensure measurability. As will be clear in the proofs (see Lemma 5.4), *any* choice of $\Omega \in \mathrm{Vor}^{(j)}_{R_0}(X_i) \cap \mathring{\mathrm{B}}_{\hat{T}_j}(\pi_{\hat{T}_j}(X_i - X_j), \rho)^c$ witnessing to the ρ-width of the Voronoi cell would lead to the same normal estimation rates as $\tilde{\eta}_i$.

As expected, when localization radii are chosen properly, Theorem 3.7 below provides quantitative bounds for boundary detection and normal estimation.

Theorem 3.7 (Guarantees for Boundary Detection and Normals)**.** *Take* $R_0 \le \frac{\tau_{\min} \wedge \tau_{\partial,\min}}{40}$. *Define*

$$r_- := \sqrt{(\tau_{\min} \wedge \tau_{\partial,\min}) R_0} \left(c_d \frac{f^5_{\max} \log n}{f^6_{\min} n R_0^d} \right)^{\frac{1}{d+1}}, r_+ := \frac{R_0}{12}, \text{ and } \rho_- := \frac{R_0}{4} =: \frac{\rho_+}{2}.$$

Then, for n *large enough, with probability at least* $1 - 4n^{-\frac{2}{d}}$, *we have that for all* $\rho \in [\rho_-, \rho_+]$ *and* $r \in [r_-, r_+]$*:*

(i) If $\partial M = \emptyset$, *then* $\mathcal{Y}_{R_0,r,\rho} = \emptyset$;

(ii) If $\partial M \neq \emptyset$ then:

(a) For all $X_i \in \mathcal{Y}_{R_0,r,\rho}$,

$$\mathrm{d}(X_i, \partial M) \leq \frac{2r^2}{\tau_{\min} \wedge \tau_{\partial,\min}};$$

(b) For all $x \in \partial M$,

$$\mathrm{d}(x, \mathcal{Y}_{R_0,r,\rho}) \leq 3r;$$

(c) For all $X_i \in \mathcal{Y}_{R_0,r,\rho}$,

$$\|\eta_{\pi_{\partial M}(X_i)} - \tilde{\eta}_i\| \leq \frac{20r}{\sqrt{R_0(\tau_{\min} \wedge \tau_{\partial,\min})}}.$$

Remark 3.8. Key quantities in Theorem 3.7 are the scale R_0 and the local bandwith r, that need to be carefully tuned in practice. Whenever prior information on the reaches $\tau_{\min}$ and $\tau_{\partial,\min}$ is at hand, we may choose R_0 as large as $\frac{\tau_{\min} \wedge \tau_{\partial,\min}}{40}$. Then, an optimal choice $r = r_-$ leads to the bounds:

(ii)a For all $X_i \in \mathcal{Y}_{R_0,r,\rho}$,

$$\mathrm{d}(X_i, \partial M) \leq (\tau_{\min} \wedge \tau_{\partial,\min}) \left(C_d \frac{f_{\max}^5}{f_{\min}^5} \frac{\log n}{n f_{\min} (\tau_{\min} \wedge \tau_{\partial,\min})^d} \right)^{\frac{2}{d+1}},$$

(ii)b For all $x \in \partial M$,

$$\mathrm{d}\left(x, \mathcal{Y}_{R_0,r,\rho}\right) \leq (\tau_{\min} \wedge \tau_{\partial,\min}) \left(C_d \frac{f_{\max}^5}{f_{\min}^5} \frac{\log n}{n f_{\min} (\tau_{\min} \wedge \tau_{\partial,\min})^d} \right)^{\frac{1}{d+1}},$$

(ii)c For all $X_i \in \mathcal{Y}_{R_0,r,\rho}$,

$$\|\eta_{\pi_{\partial M}(X_i)} - \tilde{\eta}_i\| \leq \left(C_d \frac{f_{\max}^5}{f_{\min}^5} \frac{\log n}{n f_{\min} (\tau_{\min} \wedge \tau_{\partial,\min})^d} \right)^{\frac{1}{d+1}}.$$

The proof of Theorem 3.7 is given in Section 5.1. In a nutshell, Item (i) guarantees that no false positive occur if $\partial M = \emptyset$. On the other hand, if $\partial M \neq \emptyset$, for $\varepsilon \asymp (\log n/n)^{1/(d+1)}$ and optimal choices of r_- and R_0, Items (ii)a and (ii)b ensure that $\mathcal{Y}_{R_0,r,\rho}$ is an $O(\varepsilon)$-covering of ∂M that consists of points $O(\varepsilon^2)$-close to ∂M.

In the convex case $\tau_{\min} = \infty$, taking the convex hull of $\mathcal{Y}_{R_0,r,\rho}$ — similarly to [26] — would result in an $O(\varepsilon^2)$-approximation of M, and the boundary of this convex hull in an $O(\varepsilon^2)$-approximation of ∂M. Finally, Item (ii)c asserts that the estimated normals at boundary observations are $O(\varepsilon)$-precise.

The intuition behind the respective rates $O(\varepsilon)$ and $O(\varepsilon^2)$ is the same as in the convex case of [26]: for a fixed boundary point $x \in \partial M$, the "curved rectangle" $\{u \in \mathrm{B}(x,\varepsilon) \cap M \mid \mathrm{d}(u, \partial M) \leq \varepsilon^2\}$ has volume of order $\varepsilon^{d-1} \times \varepsilon^2 = \varepsilon^{d+1}$. Hence, the choice $\varepsilon \asymp (\log n/n)^{1/(d+1)}$ ensures that these curved rectangular regions are occupied by sample points with high probability, uniformly over the choice of x on a grid. Our procedure then guarantees that these close-to-boundary sample points will be identified as such.

Remark 3.9. The above argument may be pushed further to gain insights on the number $|\mathcal{Y}_{R_0,r,\rho}|$ of detected points.

- To derive an upper bound, use Item (ii)a to get that $|\mathcal{Y}_{R_0,r,\rho}| \leq \sum_{i=1}^{n} \mathbb{1}_{\mathrm{d}(X_i,\partial M)\lesssim\varepsilon^2}$ with high probability, where $\varepsilon \asymp (\log n/n)^{1/(d+1)}$. Since $P(\{u \in M \mid \mathrm{d}(u,\partial M) \leq \varepsilon^2\}) \lesssim \mathrm{Vol}_{d-1}(\partial M)\varepsilon^2$ for ε small enough, this ensures that

$$|\mathcal{Y}_{R_0,r,\rho}| \lesssim n\varepsilon^2 \asymp \log n^{2/(d+1)} n^{(d-1)/(d+1)}.$$

  In particular, for $d = 1$, optimal choices of the parameters guarantees that the number of detected points should be no more than roughly $\log n$. In this 1-dimensional case, it falls under the intuition that the 'optimal' number of detected points should be 2, corresponding to extremal points drawn on a curve.

- On the other hand, Items (ii)a and (ii)b combined provide a lower bound on $|\mathcal{Y}_{R_0,r,\rho}|$. Letting $N(\varepsilon)$ denote the ε-covering number of ∂M, standard volume arguments show that $N(\varepsilon) \simeq \mathrm{Vol}_{d-1}(\partial M)/\varepsilon^{d-1}$. Furthermore, the fact that $\sup_{x\in\partial M} \mathrm{d}(x, \mathcal{Y}_{R_0,r,\rho}) \lesssim \varepsilon$ means that $\mathcal{Y}_{R_0,r,\rho}$ is a $O(\varepsilon)$-covering of ∂M, so that $|\mathcal{Y}_{R_0,r,\rho}| \geq N(\varepsilon)$. Hence, we obtain

$$|\mathcal{Y}_{R_0,r,\rho}| \geq N(\varepsilon) \gtrsim (\log n)^{-(d-1)/(d+1)} n^{(d-1)/(d+1)}.$$

These two matching bounds (up to $\log n$ factors) back the intuition that as d grows large, most of the mass (and hence sample) is concentrated nearby the boundary.

If no prior information on τ_M and $\tau_{\partial M}$ are available, choosing $R_0 = (\log n)^{-1}$ would meet the requirements of Theorem 3.7 for n large enough. As well, choosing $r = \sqrt{R_0 \log n}\left(\log n/(nR_0^d)\right)^{1/(d+1)}$ would asymptotically meet the requirements of Theorem 3.7. Both of these choices incur an extra $\log n$ factor in the bounds.

Still based on $\mathcal{Y}_{R_0,r,\rho}$, we extend this "hull" construction to the non-convex case by leveraging the additional tangential (Proposition 3.2) and normal (Theorem 3.7 (ii)c) estimates, to provide estimators of M and ∂M.

3.2 Boundary estimation

Assume that $\partial M \neq \emptyset$. Then ∂M is a $(d-1)$-dimensional $\mathcal{C}^2$-submanifold without boundary. Therefore, using manifold estimators of [2, 3, 39, 29] designed for the empty boundary case with input points $\mathcal{Y}_{R_0,r,\rho}$ seems relevant. We choose to focus on the manifold estimator proposed in [2], based on the Tangential Delaunay Complex [11], as it also provides a topologically consistent estimation. This procedure, as well as the aforementioned two others, takes as input boundary points but also estimates of the tangent spaces (of the boundary). Thus, a preliminary step is to provide estimators for the boundary tangent spaces at points of $\mathcal{Y}_{R_0,r,\rho}$.

Definition 3.10 (Boundary's Tangent Space Estimator)**.** For all $X_i \in \mathcal{Y}_{R_0,r,\rho}$, $\hat{T}_{\partial,i}$ is defined as the orthogonal complement of $\pi_{\hat{T}_i}(\tilde{\eta}_i)$ in $\hat{T}_i$. That is,

$$\hat{T}_{\partial,i} := (\pi_{\hat{T}_i}(\tilde{\eta}_i))^{\perp} \cap \hat{T}_i.$$

A straightforward consequence of Proposition 3.2 and Theorem 3.7 is that the estimator $\hat{T}_{\partial,i}$ is a $O\big((\log n/n)^{1/(d+1)}\big)$-approximation of $T_{\pi_{\partial M}(X_i)}\partial M$, for any $X_i \in \mathcal{Y}_{R_0,r,\rho}$.

Corollary 3.11 (Boundary's Tangent Space Estimation)**.** *Under the assumptions of Proposition 3.2 and Theorem 3.7 we have, for n large enough, with probability larger than $1 - 4n^{-\frac{2}{d}}$,*

$$\max_{X_i \in \mathcal{Y}_{R_0,r,\rho}} \angle(T_{\pi_{\partial M}(X_i)}\partial M, \hat{T}_{\partial,i}) \leq \frac{20r}{\sqrt{(\tau_{\min} \wedge \tau_{\partial,\min})R_0}}.$$

Thus, choosing $R_0 = \frac{\tau_{\min} \wedge \tau_{\partial,\min}}{40}$ and $r = r_-$ yields

$$\max_{X_i \in \mathcal{Y}_{R_0,r_-,\rho}} \angle(T_{\pi_{\partial M}(X_i)}\partial M, \hat{T}_{\partial,i}) \leq \left(C_d \frac{f_{\max}^5}{f_{\min}^5} \frac{\log n}{n f_{\min}(\tau_{\min} \wedge \tau_{\partial,\min})^d} \right)^{\frac{1}{d+1}}.$$

A short proof can be found in Appendix D.2, that connects $\angle(T_{\pi_{\partial M}(X_i)}\partial M, \hat{T}_{\partial,i})$ to $\angle(T_{X_i}M, \hat{T}_i)$ and $\angle(\eta_{\pi_{\partial M}(X_i)}, \tilde{\eta}_i)$. The estimation rate for $T_{\pi_{\partial M}(X_i)}\partial M$ is then driven by the larger of these quantities, i.e. $\angle(\eta_{\pi_{\partial M}(X_i)}, \tilde{\eta}_i)$ according to Proposition 3.2 and Theorem 3.7.

Equipped with Corollary 3.11, we are now in position to provide an estimator for ∂M. Following [2], we let $\varepsilon = C\frac{\tau_{\partial,\min}}{R_0}r$, where r and R_0 are chosen as in Theorem 3.7, and let $\mathbb{Y}_\partial$ denote an ε-sparsification of $\mathcal{Y}_{R_0,r,\rho}$, i.e. a subset of $\mathcal{Y}_{R_0,r,\rho}$ that forms an ε-covering of $\mathcal{Y}_{R_0,r,\rho}$ with ε-separated points. Such a sparsification can be obtained by running the farthest point sampling algorithm over $\mathcal{Y}_{R_0,r,\rho}$, and it results in a 2ε-covering of ∂M, according to Theorem 3.7. We also denote by $\mathbb{T}_\partial$ the collection of $\hat{T}_{\partial,i}$'s, for $X_i \in \mathbb{Y}_\partial$, and define our estimator of ∂M as the (weighted) Tangential Delaunay Complex [11] based on $(\mathbb{Y}_\partial, \mathbb{T}_\partial)$:

$$\widehat{\partial M} := \mathrm{Del}^{\omega_*}(\mathbb{Y}_\partial, \mathbb{T}_\partial).$$

Since ∂M has no boundary, [2, Theorem 4.4] applies and yields the following reconstruction result.

Theorem 3.12 (Boundary Estimation: Upper Bound)**.** *Provided that $\partial M \neq \emptyset$ and under the assumptions of Proposition 3.2 and Theorem 3.7, we have for n large enough, with probability larger than $1 - 4n^{-\frac{2}{d}}$,*

(i) $\mathrm{d_H}(\partial M, \widehat{\partial M}) \leq C_d \frac{\tau_{\partial,\min}}{R_0^2} r^2$,

(ii) ∂M and $\widehat{\partial M}$ are ambient isotopic.

As a consequence, for n large enough, choosing $R_0 = \frac{\tau_{\min} \wedge \tau_{\partial,\min}}{40}$ and $r = r_-$, we have

$$\mathbb{E}_{P^n}\left[\mathrm{d_H}(\partial M, \widehat{\partial M})\right] \leq C_d \tau_{\partial,\min} \left(\frac{f_{\max}^5}{f_{\min}^5} \frac{\log n}{n f_{\min}(\tau_{\min} \wedge \tau_{\partial,\min})^d} \right)^{\frac{2}{d+1}}.$$

The proof derives from a direct application of the reconstruction result of [2, Theorem 4.4], the assumptions of which hold with high probability, according to the distance bounds of Theorem 3.7 (ii)a and (ii)b and the angle bounds of Corollary 3.11.

Note that the ambient dimension D plays no role in Theorem 3.12, neither in the assumptions, the rate nor the constants. Interestingly, it assesses the topological correctness of our estimator $\widehat{\partial M}$, showing the particular interest of estimators based on simplicial complexes. Choosing the largest possible R_0, i.e. $R_0 = \frac{\tau_{\min} \wedge \tau_{\partial,\min}}{40}$, and $r = r_-$, Theorem 3.12 provides an upper bound on $\mathrm{d_H}(\partial M, \widehat{\partial M})$ with high probability, uniformly over the class $\mathcal{P}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}(f_{\min}, f_{\max})$ introduced in Definition 2.11. This uniform convergence rate is in line with the estimation rate $O\big((\log n/n)^{2/(d+1)}\big)$

for boundary estimation given by [45, 26], under convexity-type assumptions in the full dimensional case. Letting $\tau_{\min} = \infty$, the convex case can even be seen of as a sub-case of our class of distributions, since $\mathcal{P}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}(f_{\min}, f_{\max}) \supset \mathcal{P}^{d,D}_{\infty,\tau_{\partial,\min}}(f_{\min}, f_{\max})$. In fact, even in this simpler case, we can show that the rate $O\big((\log n/n)^{2/(d+1)}\big)$ is minimax over the class of convex submanifolds.

Theorem 3.13 (Boundary Estimation: Lower Bound)**.** *Assume that* $f_{\min} \leq c_d/\tau^d_{\partial,\min}$*, and that* $c'_d/\tau^d_{\partial,\min} \leq f_{\max}$ *for some small enough* $c_d, (c'_d)^{-1} > 0$*. Then for all* $n \geq 1$*,*

$$\inf_{\hat{B}} \sup_{P \in \mathcal{P}^{d,D}_{\infty,\tau_{\partial,\min}}(f_{\min},f_{\max})} \mathbb{E}_{P^n}\left[\mathrm{d_H}\big(\partial M, \hat{B}\big)\right] \geq C_d \tau_{\partial,\min} \left\{1 \wedge \left(\frac{1}{f_{\min}\tau^d_{\partial,\min} n}\right)^{\frac{2}{d+1}}\right\}.$$

A proof of Theorem 3.13 is given in Appendix F and relies on standard Bayesian arguments.

Since for all $\tau_{\min} > 0$, $\mathcal{P}^{d,D}_{\infty,\tau_{\partial,\min}}(f_{\min}, f_{\max}) \subset \mathcal{P}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}(f_{\min}, f_{\max})$, Theorem 3.13 and Theorem 3.16 together ensure that our boundary estimation procedure is minimax over the model $\mathcal{P}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}(f_{\min}, f_{\max})$, up to $\log n$ factors. From a statistical viewpoint, these two results show that estimating the boundary under reach conditions on M is not more difficult than estimating the boundary in the convex case.

3.3 Boundary-adaptive manifold estimation

If $\partial M = \emptyset$, it is known that M can be estimated optimally by local linear patches [3]. That is, choosing $\varepsilon_{\mathring{M}} = \left(C_d \frac{f^4_{\max} \log n}{f^5_{\min} n}\right)^{1/d}$, and estimating M via the union of tangential balls $\hat{M} = \bigcup_{i=1}^n X_i + \mathrm{B}_{\hat{T}_i}(0, \varepsilon_{\mathring{M}})$ leads to $\mathrm{d_H}(M, \hat{M}) \leq C_d f_{\max} \varepsilon^2_{\mathring{M}}/(f_{\min}\tau_{\min})$ [3, Theorem 6], recovering the minimax rate $O\big((\log n/n)^{2/d}\big)$ over the class of $\mathcal{C}^2$ manifolds without boundary [36].

If $\partial M \neq \emptyset$ and X_i is close to ∂M, a tangential ball $X_i + \mathrm{B}_{\hat{T}_i}(0, \varepsilon_{\mathring{M}})$ may go past ∂M along the normal direction $\eta_{\pi_{\partial M}(X_i)}$, leading to a poor approximation of M in terms of Hausdorff distance. In this case, replacing $X_i + \mathrm{B}_{\hat{T}_i}(0, \varepsilon_{\mathring{M}})$ by a tangential half-ball oriented at the opposite of the outward-pointing normal vector $\eta_{\pi_{\partial M}(X_i)}$ seems more appropriate. We formalize this intuition as follows.

Let $\mathcal{Y}_{R_0,r,\rho}$ denote the detected boundary observations of Definition 3.3. These points will generate half-balls, with radius $\varepsilon_{\partial M}$, that will roughly approximate the inward slab $M \cap \mathrm{B}(\partial M, \varepsilon_{\partial M})$ of radius $\varepsilon_{\partial M}$. To approximate the remaining part of M, we further define the *$\varepsilon_{\partial M}$-inner points* as

$$\mathring{\mathcal{Y}}_{\varepsilon_{\partial M}} := \{X_i \in \mathbb{X}_n \mid \mathrm{d}(X_i, \mathcal{Y}_{R_0,r,\rho}) \geq \varepsilon_{\partial M}/2\}. \tag{4}$$

Then, the manifold M may be reconstructed as follows (see Figure 6).

Definition 3.14 (Boundary-Adaptive Manifold Estimator)**.** Given some scale parameters $\varepsilon_{\mathring{M}}$ and $\varepsilon_{\partial M}$, the manifold estimator $\hat{M} := \hat{M}_{\mathrm{Int}} \cup \hat{M}_{\partial}$, is defined as

$$\hat{M}_{\mathrm{Int}} := \bigcup_{X_i \in \mathring{\mathcal{Y}}_{\varepsilon_{\partial M}}} X_i + \mathrm{B}_{\hat{T}_i}(0, \varepsilon_{\mathring{M}}),$$

$$\hat{M}_{\partial} := \bigcup_{X_i \in \mathcal{Y}_{R_0,r,\rho}} \Big(X_i + \mathrm{B}_{\hat{T}_i}(0, \varepsilon_{\partial M})\Big) \cap \{z, \langle z - X_i, \tilde{\eta}_i\rangle \leq 0\},$$

with

- the $\hat{T}_i$'s being the estimated tangent spaces from Proposition 3.2,

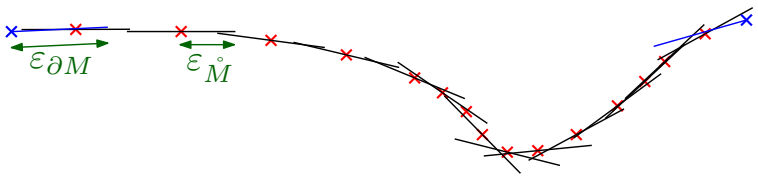


Figure 6: The local linear estimator $\hat{M}$ from Definition 3.14 for $d = 1$ and $D = 2$. The boundary estimator $\hat{M}_\partial$ corresponds to the union of the two blue segments, and $\hat{M}_{\mathrm{Int}}$ to that of the black segments.

- the $\tilde{\eta}_i$'s being the estimated of the outward-pointing normals from Theorem 3.7.

Note that $\hat{M}$ is adaptive in the sense that it does not require information about emptiness of ∂M. If $\partial M = \emptyset$, then $\mathcal{Y}_{R_0,r,\rho} = \emptyset$ with high probability (Theorem 3.7 (i)). In this case $\hat{M}$ coincides (with high probability) with the estimator from [3], which is minimax over the class of boundariless $\mathcal{C}^2$-manifolds. Theorem 3.15 below extends the error bound for $\hat{M}$ whenever $\partial M \neq \emptyset$.

Theorem 3.15 (Estimation with Boundary: Upper Bound)**.** *Choose (R_0, r, ρ) as in Theorem 3.7, set*

$$\varepsilon_{\mathring{M}} = \left(C_d \frac{\log n}{f_{\min} n}\right)^{\frac{1}{d}} \text{ and } \varepsilon_{\partial M} = 18r.$$

Then for n large enough, with probability larger than $1 - 4n^{-\frac{2}{d}}$, we have

$$\mathrm{d_H}(M, \hat{M}) \leq C_d \begin{cases} (f_{\max}/f_{\min})^{\frac{4}{d}+1} \varepsilon^2_{\mathring{M}}/\tau_{\min} & \text{if } \partial M = \emptyset, \\ \varepsilon^2_{\partial M}/R_0 & \text{if } \partial M \neq \emptyset. \end{cases}$$

As a consequence, for n large enough, with $R_0 = \frac{\tau_{\min} \wedge \tau_{\partial,\min}}{40}$ and $r = r_-$, it holds

$$\mathbb{E}_{P^n}\left[\mathrm{d_H}(M, \hat{M})\right] \leq C_d \begin{cases} \tau_{\min} \left(\dfrac{f_{\max}^{2+d/2}}{f_{\min}^{2+d/2}} \dfrac{\log n}{f_{\min} \tau_{\min}^d n}\right)^{\frac{2}{d}} & \text{if } \partial M = \emptyset, \\ (\tau_{\min} \wedge \tau_{\partial,\min}) \left(\dfrac{f_{\max}^5}{f_{\min}^5} \dfrac{\log n}{f_{\min} (\tau_{\min} \wedge \tau_{\partial,\min})^d n}\right)^{\frac{2}{d+1}} & \text{if } \partial M \neq \emptyset. \end{cases}$$

A proof of Theorem 3.15 is given in Section 5.4. Again, note that Theorem 3.15 is completely oblivious to the ambient dimension D. In the empty boundary case, $\hat{M}$ achieves the rate $O\big((\log n/n)^{2/d}\big)$, which is minimax [36]. Whenever ∂M is not empty, the given convergence rate of $\hat{M}$ coincides with that of $\widehat{\partial M}$ for boundary estimation (Theorem 3.12), as well as that of [26, Corollary 1] for convex domains, and that of [45, Theorem 3] for r-convex domains. Note that these last two convexity-type assumptions are stronger than the bounded reach assumption for M and ∂M, so that Theorem 3.15 generalizes [26, 45]. As for the boundary estimation problem, we show that this rate $O\big((\log n/n)^{2/d}\big)$ is in fact minimax optimal over the class of d-dimensional convex domains (i.e. $\tau_{\min} = \infty$), up to $\log n$ factors.

Theorem 3.16 (Manifold Estimation: Lower Bounds)**.**

(Boundaryless) *Assume that* $f_{\min} \leq c_d/\tau_{\min}^d$ *and that* $c'_d/\tau_{\min}^d \leq f_{\max}$*, for some small enough* $c_d, (c'_d)^{-1} > 0$*. If* $d \leq D-1$*, then for all* $n \geq 1$*,*

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}^{d,D}_{\tau_{\min},\infty}(f_{\min},f_{\max})} \mathbb{E}_{P^n}\left[\mathrm{d_H}\big(M,\hat{M}\big)\right] \geq C_d \tau_{\min} \left\{1 \wedge \left(\frac{1}{f_{\min}\tau_{\min}^d n}\right)^{\frac{2}{d}}\right\}.$$

(Convex) *Assume that* $f_{\min} \leq c_d/\tau_{\partial,\min}^d$ *and* $c'_d/\tau_{\partial,\min}^d \leq f_{\max}$*, for some small enough* $c_d, (c'_d)^{-1} > 0$*. Then for all* $n \geq 1$*,*

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}^{d,D}_{\infty,\tau_{\partial,\min}}(f_{\min},f_{\max})} \mathbb{E}_{P^n}\left[\mathrm{d_H}\big(M,\hat{M}\big)\right] \geq C_d \tau_{\partial,\min} \left\{1 \wedge \left(\frac{1}{f_{\min}\tau_{\partial,\min}^d n}\right)^{\frac{2}{d+1}}\right\}.$$

The proof of Theorem 3.16 relies on the same bayesian arguments as Theorem 3.13 (see Appendix F). The first point is a slight refinement of the $\mathcal{C}^2$ case of [3, Theorem 7], as it exhibits the dependency on $\tau_{\min}$ and $f_{\min}$ of the minimax rates over the class of $\mathcal{C}^2$ manifolds without boundary. Note also that in this case, the assumption $d \leq D-1$ clearly is necessary for the model not to be empty.

Interestingly, this shows that the upper bound given in Theorem 3.15 for the empty boundary case is sharp with respect to $\tau_{\min}$. The second point of Theorem 3.16 provides the minimax rate for manifold estimation over the class of convex domains whose boundary has bounded reach. In terms of sample size, this shows that our estimator has the best possible convergence rate $O\big((\log n/n)^{2/(d+1)}\big)$ (up to $\log n$ factors) in the convex case, as well as the two procedures of [26, 45]. As for the boundary estimation problem, this result intuitively carries the message that estimating a manifold with boundary under reach conditions is not more difficult than estimating a d-dimensional convex $\mathcal{C}^2$-domain. In other words, for $\partial M \neq \emptyset$ and a fixed boundary's convexity radius $\tau_{\partial,\min}$, no additional gain can be expected from requiring a large convexity radius for the manifold (driven by $\tau_{\min}$). At last, Theorem 3.15 shows that the given dependency on the reach boundary $\tau_{\partial,\min}$ is sharp, at least in the case where $\tau_{\partial,\min} \leq \tau_{\min}$. Whether the tradeoff between $\tau_{\min}$ and $\tau_{\partial,\min}$ exhibited in Theorem 3.15 is sharp in general remains an open question.

4 Conclusion and further perspectives

Both generalizing over full dimensional $\mathcal{C}^2$ domains and boundaryless $\mathcal{C}^2$-submanifolds, this work derives nearly tight minimax upper and lower bounds for $\mathcal{C}^2$-submanifold estimation with possibly non-empty $\mathcal{C}^2$ boundary. Both the boundary estimator and the manifold estimator exhibit rates that are independent of the ambient dimension, which is of critical interest in the regime $d \ll D$ to achieve efficient dimensionality reduction. To our knowledge, this is the first instance of a statistical study dealing with general submanifold with boundary.

This work is the first minimax estimation study on manifolds with boundary. Hence, the focus has not been put on computational aspects. Yet, the proposed method is fully constructive and can easily be implemented using PCA and computational geometric algorithms. Given the space constraints, we refer the interested reader to Section 6, which discusses computational complexity, parameter tuning, and provides a few numerical examples.

On the geometric side, a significant further direction of research pertains to manifold estimation with boundary in smoother models than $\mathcal{C}^2$, such as those introduced in [3]. Beyond Hausdorff minimax optimality, an interesting feature of the boundary estimator of Theorem 3.12 is its topological exactness. This property is made possible by the fact that $\partial(\partial M) = \emptyset$ and the existence of

constructive triangulations that reconstruct boundaryless submanifolds (see [2, Theorem 4.4]). In contrast, topologically exact reconstruction methods of manifolds with boundary are only known in the specific case of *isomanifolds* (see [14, Theorem 43]), which led us to stick to an unstructured estimator with linear patches in this case (see Theorem 3.15).

On the statistical side, a major limitation of this work is the absence of noise. The proposed method would exhibit the same rates if noise of amplitude $\sigma \ll (\log n/n)^{2/d}\mathbb{1}_{\partial M=\emptyset} + (\log n/n)^{2/(d+1)}\mathbb{1}_{\partial M\neq\emptyset}$ is added, but it is likely to fail otherwise as it is based on the data points themselves. Such instabilities are common in the geometric inference literature [20, 1, 9, 23], and noise is often assumed to vanish as n goes to ∞. However, a recent line of works in the boundariless case exhibited various iterative denoising procedures that tend to relax this assumption. See for instance [29, 43, 7]. Whether such algorithms could be adapted for $\partial M \neq \emptyset$ is of particular interest.

Acknowledgments

We are grateful to the members of the *Laboratoire de Probabilités, Statistique et Modélisation* and the *Laboratoire de Mathématiques Blaise Pascal* for their insightful comments.

5 Proofs outline

Due to space constraints, the geometric results necessary to the proofs given below are deferred to the appendix (Appendix A).

5.1 Proof of Theorem 3.7

The main boundary detection result is based on the following geometric and purely deterministic result.

Theorem 5.1 (Deterministic Layout for Boundary Detection and Normals)**.** *Let*

$$R_0 \leq \frac{\tau_{\min}}{32}, r_0 \leq \frac{R_0 \wedge \tau_{\partial,\min}}{4}, r \leq \frac{R_0}{12},$$
$$\theta \leq \frac{1}{24}, \varepsilon_1 \leq \frac{r}{4}, \varepsilon_2 \leq \frac{r_0}{120} \wedge \frac{r^2}{\tau_{\min} \wedge \tau_{\partial,\min}},$$
$$\text{and } 3r \leq \rho_- < \rho_+ \leq \frac{\tau_{\min} \wedge \tau_{\partial,\min}}{80}.$$

Assume that we have:

1. *A point cloud $\mathcal{X}_n \subset M$ such that $\mathrm{d_H}(M, \mathcal{X}_n) \leq \varepsilon_1$,*
2. *Estimated tangent spaces T_j such that $\max_{1\leq j\leq n} \angle(T_{X_j}M, T_j) \leq \theta$.*

For $x \in \partial M$ and $j \in \{1, \ldots, n\}$ such that $\angle(T_xM, T_j) < 1$, write $\eta_j^*(x)$ for the unit vector of $Nor(x, M) \cap T_j$ (see Proposition A.5). Defining $\mathcal{Y}_j := \pi_{T_j}(\mathrm{B}(X_j, R_0) \cap \mathcal{X}_n - X_j)$, assume furthermore that:

3. *For all $x \in \partial M$ and $X_j \in \mathcal{X}_n \cap \mathrm{B}(x, 2r)$, for all $\rho \geq \rho_-$ and $\Omega \in T_j$ such that $\|\Omega - (\pi_{T_j}(x - X_j) - r_0\eta_j^*(x))\| \leq r_0 + \rho - \varepsilon_2$ we have $\mathrm{B}(\Omega, \rho) \cap \mathcal{Y}_j \neq \emptyset$.*

Then for all $\rho \in [\rho_-, \rho_+]$, using notation of Definitions 3.3 and 3.5, the following holds:

(i) If $\partial M = \emptyset$, then $\mathcal{Y}_{R_0,r,\rho} = \emptyset$.

(ii) If $\partial M \neq \emptyset$, then,

(a) For all $X_i \in \mathcal{Y}_{R_0,r,\rho}$,

$$\mathrm{d}(X_i, \partial M) \leq \frac{2r^2}{\tau_{\min} \wedge \tau_{\partial,\min}},$$

(b) For all $x \in \partial M$,

$$\mathrm{d}(x, \mathcal{Y}_{R_0,r,\rho}) \leq 3r.$$

(c) For all $X_i \in \mathcal{Y}_{R_0,r,\rho}$ with associated $X_j \in J_{R_0,r,\rho}(X_i)$ and witness $\Omega \in \mathrm{Vor}_{R_0}^{(j)}(X_i) \cap \mathring{\mathrm{B}}_{T_j}(\pi_{T_j}(X_i - X_j), \rho)^c$,

$$\|\eta_{\pi_{\partial M}(X_i)} - \tilde{\eta}_i^{(j)}\| \leq 4\theta + 8\sqrt{\frac{\tau_{\min} \wedge \tau_{\partial,\min}}{\rho \wedge r_0}} \frac{r}{\tau_{\min} \wedge \tau_{\partial,\min}},$$

where $\tilde{\eta}_i^{(j)} = \big(\Omega - \pi_{T_j}(X_i - X_j)\big) \,/\, \big\|\Omega - \pi_{T_j}(X_i - X_j)\big\|$.

A proof of Theorem 5.1 is given in the following Section 5.2. Figure 7 below illustrates the role of the different parameters involved in this result. In the assumptions, Items 1 and 2 require that M is sampled densely enough, and that the tangent spaces at sample points have been estimated with precision θ. In light of Proposition 3.2, these assumptions will be satisfied at scales $\varepsilon_1, \theta = O((\log n/n)^{1/d})$ for a random n-sample and a standard tangent space estimator, with high probability. Then, Item 3 basically requires that the "curved rectangles" $\{u \in \mathrm{B}(x, \sqrt{\varepsilon_2}) \cap M \mid \mathrm{d}(u, \partial M) \leq \varepsilon_2\}$ nearby all $x \in \partial M$ are occupied by sample points (see Figure 7). This assumption is key for identifying boundary observations and estimating normals accurately. The volume heuristic given below Theorem 3.7 suggests that for a random n-sample, this assumption is satisfied with high probability at scale $\varepsilon_2 = O((\log n/n)^{2/(d+1)})$.

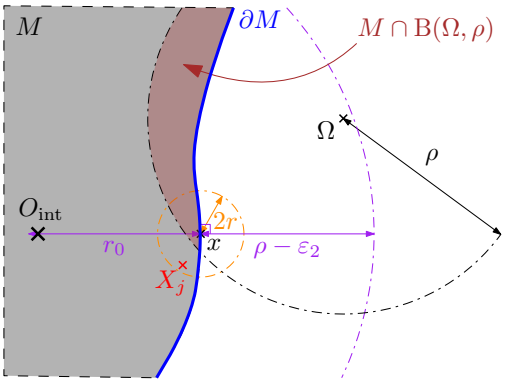


Figure 7: Illustration of Item 3 of Theorem 5.1 in full dimension $d = D = 2$, yielding the simplification that $\pi_{T_j} = \mathrm{Id}_d$ for all $j \in \{1, \ldots, n\}$. Here, we denoted $O_{\mathrm{int}} := \pi_{T_j}(x - X_j) - r_0 \eta_j^*(x)$. The assumption $\mathrm{B}(\Omega, \rho) \cap \mathcal{Y}_j \neq \emptyset$ means that the (brown) zone $\mathrm{B}(\Omega, \rho) \cap M$ contains sample points. In full generality, when $d < D$, a similar layout can be drawn, with $M, \partial M$ and $M \cap \mathrm{B}(\Omega, \rho)$ replaced by $\pi_{T_j}(M), \pi_{T_j}(\partial M)$ and $\pi_{T_j}(M) \cap \mathrm{B}(\Omega, \rho)$.

The precise statement ensuring that the conditions of Theorem 5.1 are fulfilled with high probability for a random n-sample goes as follows.

Proposition 5.2. *Fix* $R_0 \leq \frac{\tau_{\min} \wedge \tau_{\partial,\min}}{40}$, *define* $\rho_- = r_0 = \frac{R_0}{4}$, $\rho_+ = \frac{R_0}{2}$, *and set*

$$\varepsilon_2 = r_0 \left(C_d \frac{f_{\max}^5}{f_{\min}^5} \frac{\log n}{f_{\min}(n-1) r_0^d} \right)^{\frac{2}{d+1}}.$$

Then for n *large enough, the following statements hold with probability larger than* $1 - 3n^{-\frac{2}{d}}$*: for all* $i \in \{1, \ldots, n\}$,

(i) $\angle(T_{X_i} M, \hat{T}_i) \leq \dfrac{1}{\tau_{\min}} \left(C_d \dfrac{f_{\max}^{4+d}}{f_{\min}^{5+d}} \dfrac{\log n}{n} \right)^{\frac{1}{d}} \leq 1/24$;

(ii) for all $(x, \Omega) \in (\mathrm{B}(X_i, r_0) \cap \partial M) \times \hat{T}_i$,

$$\|\Omega - (\pi_{\hat{T}_i}(x - X_i) - r_0 \eta_i^*(x))\| \leq r_0 + \rho - \varepsilon_2 \quad \Rightarrow \quad \mathrm{B}(\Omega, \rho) \cap \mathcal{Y}_i \neq \emptyset,$$

where $\eta_i^*(x)$ *denotes the unique unit vector of* $Nor(x, M) \cap \hat{T}_i$ *(see Proposition A.5) .*

A proof of Proposition 5.2 is given in Section 5.3.

Proof of Theorem 3.7. Combining Proposition 5.2 and Lemma A.9 ensures that the requirements of Theorem 5.1 are fulfilled, with probability larger than $1 - 4n^{-2/d}$ for n large enough, by choosing $T_i = \hat{T}_i$ and the following set-up:

$$R_0 \leq \frac{\tau_{\min} \wedge \tau_{\partial,\min}}{40} \quad , \quad \frac{R_0}{2} = \rho_+ \geq \rho \geq \rho_- = r_0 = \frac{R_0}{4},$$

$$\varepsilon_1 = \left(C_d \frac{\log n}{f_{\min} n} \right)^{\frac{1}{d}} \quad , \quad \varepsilon_2 = r_0 \left(C_d \frac{f_{\max}^5}{f_{\min}^5} \frac{\log n}{f_{\min} n r_0^d} \right)^{\frac{2}{d+1}},$$

$$\sqrt{(\tau_{\min} \wedge \tau_{\partial,\min}) \varepsilon_2} = r_- \leq r \leq r_+ = \frac{R_0}{12} \quad , \quad \theta = \frac{1}{\tau_{\min}} \left(C_d \frac{f_{\max}^{4+d}}{f_{\min}^{5+d}} \frac{\log n}{n} \right)^{\frac{1}{d}} \leq \frac{r_-}{\tau_{\min} \wedge \tau_{\partial,\min}}. \quad \square$$

5.2 Proof of Theorem 5.1

We decompose the proof into three intermediate results. As a first step, we prove that the sample points witnessing for boundary observations — i.e. points X_j making $J_{R_0,r,\rho}(X_i)$ nonempty, see (2) —, must be close to ∂M. In fact, we show that they must be among the points X_j's on which the Assumption 3 of Theorem 5.1 holds. See Appendix C for the proof.

Lemma 5.3. *Under the assumptions of Theorem 5.1, if* $X_j \in J_{R_0,r,\rho}(X_i)$, *then* $\partial M \neq \emptyset$ *and*

$$\mathrm{d}(X_j, \partial M) \leq 2r.$$

The next step builds upon Lemma 5.3, to guarantee that the detected boundary observations — i.e. points X_i such that $J_{R_0,r,\rho}(X_i) \neq \emptyset$ — are close to the boundary ∂M, and that the associated estimated normals are close to the true normals at boundary points. In other words, we prove Theorem 5.1 (ii)a and (ii)c.

Lemma 5.4 (Theorem 5.1 (ii)a and (ii)c)**.** *Under the assumptions of Theorem 5.1, for all* $X_i \in \mathcal{Y}_{R_0,r,\rho}$,

$$\mathrm{d}(X_i, \partial M) \leq 2\varepsilon_2,$$

and for all witness $X_j \in J_{R_0,r,\rho}(X_i)$,

$$\|\eta_{\pi_{\partial M}(X_i)} - \tilde{\eta}_i^{(j)}\| \leq 4\left(\theta + \sqrt{\left(\frac{1}{\rho} + \frac{1}{r_0}\right)\varepsilon_2} + \frac{4r}{\tau_{\min}}\right).$$

Proof of Lemma 5.4. To begin with, note that as $X_i \in \mathcal{Y}_{R_0,r,\rho}$ has some witness $X_j \in J_{R_0,r,\rho}(X_i)$, Lemma 5.3 entails that $\partial M \neq \emptyset$. Also, since $\|X_i - X_j\| \leq r \leq \tau_{\min}/48$, Proposition A.2 and Lemma A.1 yield that

$$\angle(T_{X_i}M, T_j) \leq \theta + \frac{2r}{\tau_{\min}} \leq \frac{1}{24} + \frac{1}{24} \leq \frac{1}{12}. \tag{5}$$

Furthermore, Lemma 5.3 and triangle inequality gives

$$\mathrm{d}(X_i, \partial M) \leq \|X_i - X_j\| + \mathrm{d}(X_j, \partial M) \leq 3r,$$

so that $x' := \pi_{\partial M}(X_i) \in \partial M$ satisfies $\|x' - X_i\| \leq 3r \leq R_0$.

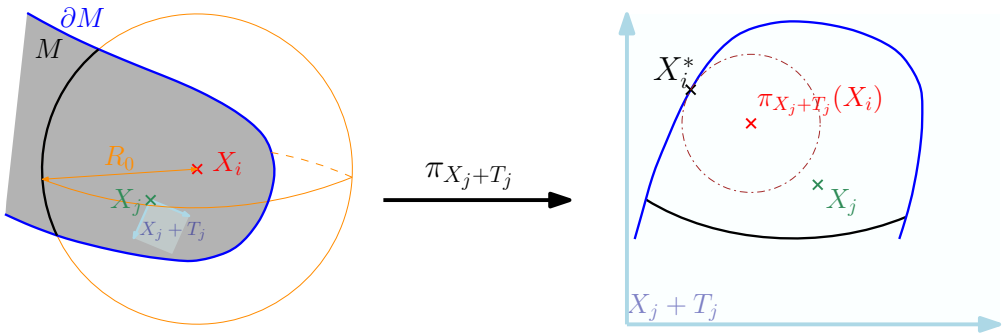


Figure 8: Layout for the proof of Lemma 5.4.

Consider $X_i^* \in \operatorname{argmin}_{z \in \pi_{X_j+T_j}(\partial M \cap \mathrm{B}(X_i,R_0))} \|z - \pi_{X_j+T_j}(X_i)\|$ (see Figure 8). As $x' \in \partial M \cap \mathrm{B}(X_i, R_0)$, $\pi_{X_j+T_j}(x')$ lies in the set where the argmin defining X_i^* ranges, and hence

$$\left\|X_i^* - \pi_{X_j+T_j}(X_i)\right\| \leq \left\|\pi_{X_j+T_j}(x') - \pi_{X_j+T_j}(X_i)\right\| \leq \left\|x' - X_i\right\| \leq 3r. \tag{6}$$

Introduce now $x \in \partial M \cap \mathrm{B}(X_i, R_0)$ such that $\pi_{T_j}(x - X_j) = X_i^*$. From $\|x - X_i\| \leq R_0$ only, Proposition A.4, (5) and (6) actually guarantee that

$$\|x - X_i\| \leq \frac{\left\|X_i^* - \pi_{X_j+T_j}(X_i)\right\|}{1 - \frac{1}{12} - \frac{\|x - X_i\|}{2\tau_{\min}}} \leq 4r.$$

Applying Proposition A.2 and Proposition B.1 yields that

$$\begin{aligned}\angle(T_xM, T_j) &\leq \angle(T_xM, T_{X_i}M) + \angle(T_{X_i}M, T_{X_j}M) + \angle(T_{X_i}M, T_j) \\ &\leq \frac{2\|X_i - x\|}{\tau_{\min}} + \frac{2\|X_j - X_i\|}{\tau_{\min}} + \theta \\ &\leq \frac{10r}{\tau_{\min}} + \theta. \end{aligned} \tag{7}$$

In particular, $\angle(T_xM, T_j) \leq 1/8$, so that Corollary A.8 asserts that

$$\pi_{X_j+T_j}(X_i) - X_i^* = -\left\|X_i^* - \pi_{X_j+T_j}(X_i)\right\| \eta_j^*(x),$$

where $\eta_j^*(x)$ is *the* unit vector of $T_j \cap Nor(x, M)$ (see Proposition B.7). Now, we write $O := X_i^* - r_0\eta_j^*(x)$. Recall that by definition, since $X_j \in J_{R_0,r,\rho}(X_i)$, there exists $\Omega = \pi_{T_j}(X_i - X_j) + \rho\tilde{\eta}_i^{(j)} \in T_j$ such that $\mathrm{B}(\Omega, \rho) \cap \mathcal{Y}_j = \emptyset$.

On one hand, since $\mathrm{B}(\Omega, \rho) \cap \mathcal{Y}_j = \emptyset$, Assumption 3 of Theorem 5.1 implies that $\|\Omega - O\| \geq r_0 + \rho - \varepsilon_2$. On the other hand, we can develop

$$\begin{aligned}
\|\Omega - O\| &= \left\|\big(r_0 - \left\|X_i^* - \pi_{X_j+T_j}(X_i)\right\|\big)\eta_j^*(x) + \rho\tilde{\eta}_i^{(j)}\right\| \\
&= \sqrt{\big(\rho + r_0 - \left\|X_i^* - \pi_{X_j+T_j}(X_i)\right\|\big)^2 - 2\rho\big(r_0 - \left\|X_i^* - \pi_{X_j+T_j}(X_i)\right\|\big)(1 - \langle\eta_j^*(x), \tilde{\eta}_i^{(j)}\rangle)} \\
&\leq \rho + r_0 - \left\|X_i^* - \pi_{X_j+T_j}(X_i)\right\| - \frac{\rho\big(r_0 - \left\|X_i^* - \pi_{X_j+T_j}(X_i)\right\|\big)(1 - \langle\eta_j^*(x), \tilde{\eta}_i^{(j)}\rangle)}{\rho + r_0 - \left\|X_i^* - \pi_{X_j+T_j}(X_i)\right\|}.
\end{aligned}$$

Hence, combining the two above bounds on $\|\Omega - O\|$ solves to

$$\left\|X_i^* - \pi_{X_j+T_j}(X_i)\right\| + \frac{\rho\big(r_0 - \left\|X_i^* - \pi_{X_j+T_j}(X_i)\right\|\big)(1 - \langle\eta_j^*(x), \tilde{\eta}_i^{(j)}\rangle)}{\rho + r_0 - \left\|X_i^* - \pi_{X_j+T_j}(X_i)\right\|} \leq \varepsilon_2. \tag{8}$$

From Equation (8), we can now conclude readily.

- To bound $\mathrm{d}(X_i, \partial M)$, note that (8) gives $\left\|X_i^* - \pi_{X_j+T_j}(X_i)\right\| \leq \varepsilon_2$. Therefore, Proposition A.4 yields Theorem 5.1 (ii)a by writing

$$\mathrm{d}(X_i, \partial M) \leq \|X_i - x\| \leq \frac{\left\|X_i^* - \pi_{X_j+T_j}(X_i)\right\|}{1 - \angle(T_xM, T_j) - \|X_i - x\|/(2\tau_{\min})} \leq 2\varepsilon_2.$$

- To bound $\|\eta_{\pi_{\partial M}(X_i)} - \tilde{\eta}_i^{(j)}\|$, note that (8) and the fact that $r_0 \geq 2\varepsilon_2$ also yield

$$1 - \langle\eta_j^*(x), \tilde{\eta}_i^{(j)}\rangle \leq \frac{\rho + r_0}{\rho(r_0 - \varepsilon_2)}\varepsilon_2 \leq 2\left(\frac{1}{\rho} + \frac{1}{r_0}\right)\varepsilon_2.$$

  As $\eta_j^*(x)$ and $\tilde{\eta}_i^{(j)}$ are both unit vectors, this leads to

$$\|\eta_j^*(x) - \tilde{\eta}_i^{(j)}\| = \sqrt{2(1 - \langle\eta_j^*(x), \tilde{\eta}_i^{(j)}\rangle)} \leq 2\sqrt{\left(\frac{1}{\rho} + \frac{1}{r_0}\right)\varepsilon_2}. \tag{9}$$

  In addition, Proposition B.8 and bound (7) combine to

$$\|\eta_j^*(x) - \eta_x\| \leq \sqrt{2}\angle(T_xM, T_j) \leq \sqrt{2}\left(\frac{10r}{\tau_{\min}} + \theta\right). \tag{10}$$

  Finally, triangle inequality yields

$$\|x - \pi_{\partial M}(X_i)\| \leq \|X_i - x\| + \mathrm{d}(X_i, \partial M) \leq 4\varepsilon_2 \leq (\tau_{\min} \wedge \tau_{\partial,\min})/32,$$

  so that Proposition A.3 asserts that

$$\|\eta_{\pi_{\partial M}(X_i)} - \eta_x\| \leq \frac{36}{\tau_{\min} \wedge \tau_{\partial,\min}}\varepsilon_2 \leq 2\sqrt{\left(\frac{1}{\rho} + \frac{1}{r_0}\right)\varepsilon_2}. \tag{11}$$

  Combining Equations (9) to (11) with triangle inequality concludes the proof of Theorem 5.1 (ii)c and that of Lemma 5.4. □

The last point (ii)b of Theorem 5.1 derives from the following lemma.

Lemma 5.5 (Theorem 5.1 (ii)b)**.** *Under the assumptions of Theorem 5.1, if $\partial M \neq \emptyset$, then for all $x \in \partial M$, there exists $X_i \in \mathcal{Y}_{R_0,r,\rho}$ such that*

$$\mathrm{d}(x, \mathcal{Y}_{R_0,r,\rho}) \leq 3r.$$

Proof of Lemma 5.5. Let $x \in \partial M$, and assume without loss of generality that we have $\|x - X_1\| = \min_{1\leq i\leq n} \|x - X_i\|$. We thus have $\|x - X_1\| \leq \varepsilon_1 \leq R_0$. Similarly to the proof of Lemma 5.4, define

$$X_1^* \in \operatorname*{argmin}_{z\in\pi_{X_1+T_1}(\partial M\cap \mathrm{B}(X_1,R_0))} \|z - X_1\|,$$

and take $y \in \partial M \cap \mathrm{B}(X_1, R_0)$ such that $\pi_{X_1+T_1}(y) = X_1^*$. As $x \in \partial M \cap \mathrm{B}(X_1, R_0)$, we have $\|X_1^* - X_1\| \leq \|\pi_{X_1+T_1}(x - X_1)\| \leq \|x - X_1\| \leq \varepsilon_1$, so that Proposition A.4 entails

$$\|y - X_1\| \leq \frac{\varepsilon_1}{1 - \theta - \frac{R_0}{2\tau_{\min}}} \leq 2\varepsilon_1.$$

Since $\theta \leq 1/24$ and $\varepsilon_1 \leq \tau_{\min}/120$, Propositions B.1 and A.2 yield that

$$\angle(T_yM, T_1) \leq \angle(T_yM, T_{X_1}M) + \angle(T_{X_1}M, T_1) \leq \frac{2\,\|X_1 - y\|}{\tau_{\min}} + \theta \leq 1/8.$$

Hence, let $\eta_1^*(y)$ be *the* unit vector of $Nor(y, M) \cap T_1$ (see Proposition B.7). In turn, Lemma A.7 applied at y asserts that

$$\mathring{\mathrm{B}}_{y+T_1}\big(y + 2\rho_+\eta_1^*(y), 2\rho_+\big) \cap \pi_{y+T_1}(\mathrm{B}(y, \tau_{\min}/16) \cap M) = \emptyset.$$

Since $R_0 \leq \tau_{\min}/32$, $\mathrm{B}(X_1, R_0) \subset \mathrm{B}(y, \tau_{\min}/16)$. Moreover, $\pi_{X_1+T_1}(\mathrm{B}(X_1, R_0) \cap M) = (X_1 - y)^\perp + \pi_{y+T_1}(\mathrm{B}(X_1, R_0) \cap M)$ and $(X_1 - y)^\perp = (X_1^* - y)^\perp$, and hence

$$\mathring{\mathrm{B}}_{X_1+T_1}(X_1^* + 2\rho_+\eta_1^*(y), 2\rho_+) \cap \pi_{X_1+T_1}(\mathrm{B}(X_1, R_0) \cap M) = \emptyset. \tag{12}$$

Since $\rho \leq 2\rho_+$, we deduce that $\mathring{\mathrm{B}}_{X_1+T_1}(X_1^* + \rho\eta_1^*(y), \rho) \cap (X_1 + \mathcal{Y}_1) = \emptyset$. Now, consider

$$\delta := \min\left\{t > 0, \mathrm{B}_{X_1+T_1}\big(X_1^* + (\rho - t)\eta_1^*(y), \rho\big) \cap (X_1 + \mathcal{Y}_1) \neq \emptyset\right\},$$

Since for all $t \geq \varepsilon_2$, the point $\Omega_t := X_1^* - X_1 + (\rho - t)\eta_1^*(y) \in T_1$ satisfies

$$\|\Omega_t - (\pi_{T_1}(y - X_1) - r_0\eta_1^*(y))\| = r_0 + \rho - t \leq r_0 + \rho - \varepsilon_2,$$

Assumption 3 of Theorem 5.1 forces to have $\mathrm{B}(\Omega_t, \rho) \cap \mathcal{Y}_1 \neq \emptyset$, and hence $\delta \leq \varepsilon_2$.

By construction of δ, there exists $z = \pi_{X_1+T_1}(X_{i_0}) \in \partial\mathrm{B}_{X_1+T_1}(X_1^* + (\rho - \delta)\eta_1^*(y), \rho) \cap (X_1 + \mathcal{Y}_1)$. We may decompose z as $z = X_1^* + \alpha\eta_1^*(y) + \beta v$, where v is a unit vector of $T_1 \cap \mathrm{span}(\eta_1^*(y))^\perp$. Since $z \in \partial\mathrm{B}_{X_1+T_1}(X_1^* + (\rho - \delta)\eta_1^*(y), \rho)$ and $z \in (X_1 + \mathcal{Y}_1) \subset \mathring{\mathrm{B}}_{X_1+T_1}(X_1^* + 2\rho_+\eta_1^*(y), 2\rho_+)^c$ from (12), we have

- $\|z - (X_1^* + (\rho - \delta)\eta_1^*(y))\| = \rho$, and thus $(\alpha - \rho + \delta)^2 + \beta^2 = \rho^2$;
- $\|z - (X_1^* + 2\rho_+\eta_1^*(y))\| \geq 2\rho_+$, and thus $(\alpha - 2\rho_+)^2 + \beta^2 \geq 4\rho_+^2$.

Therefore, after developing the above, we get that $\|X_1^* - z\|^2 = \alpha^2 + \beta^2$ satisfies

$$\begin{cases} \|X_1^* - z\|^2 = 2\rho\delta - \delta^2 + 2\alpha(\rho - \delta) \leq 2\rho\delta + 2\rho_+\alpha, \\ \|X_1^* - z\|^2 \geq 4\rho_+\alpha = 2(2\rho_+\alpha), \end{cases}$$

which yields $\|X_1^* - z\|^2 \leq 4\rho\delta \leq 4\rho\varepsilon_2$.

Also by construction, we have $\mathring{\mathrm{B}}_{T_1}(\Omega_\delta, \rho) \cap \mathcal{Y}_1 = \emptyset$ and $\|\Omega_\delta - \pi_{X_1+T_1}(X_{i_0} - X_1)\| = \|\Omega_\delta - z\| = \rho$ As a result, it is clear that if $\|X_{i_0} - X_1\| \leq r$, then $\Omega_\delta \in \mathrm{Vor}_{R_0}^{(1)}(X_{i_0})$, which yields $X_1 \in J_{R_0,\rho,r}(X_{i_0})$ and hence $X_i \in \mathcal{Y}_{R_0,r,\rho}$. Therefore, it remains to prove that $\|X_{i_0} - X_1\| \leq r$ to conclude the proof. For this, simply write

$$\|\pi_{X_1+T_1}(X_i) - X_1\| \leq \|z - X_1^*\| + \|X_1^* - X_1\| \leq (\varepsilon_1 + 2\sqrt{\rho\varepsilon_2}),$$

and since $X_{i_0} \in \mathrm{B}(X_1, R_0)$, Proposition A.4 applied at X_1 yields

$$\|X_i - X_1\| \leq 2\,(\varepsilon_1 + 2\sqrt{\rho\varepsilon_2}) \leq r.$$

As a result, we can conclude the proof of Lemma 5.5 (and Theorem 5.1 (ii)b) by noting that

$$\mathrm{d}(x, \mathcal{Y}_{R_0,\rho,r}) \leq \|x - X_{i_0}\| \leq \|x - X_1\| + \|X_1 - X_{i_0}\| \leq 3r. \qquad \square$$

5.3 Proof of Proposition 5.2

Proof of Proposition 5.2. Without loss of generality we fix $i = 1$, and work conditionally on X_1. Let A_1 denote the event

$$A_1 := \left\{ \angle(T_{X_1}M, \hat{T}_1) \leq C_d \left(\frac{f_{\max}^{4+d} \log n}{f_{\min}^{5+d} \tau_{\min}^d (n-1)} \right)^{1/d} \right\},$$

which has probability larger than $1 - 2\,(1/n)^{1+\frac{2}{d}}$ from Proposition 3.2. Note that A_1 is $\sigma(Y_2, \ldots, Y_n)$-measurable, where $Y_i = X_i \mathbb{1}_{X_i \in \mathrm{B}(X_1,h)}$. We further assume that n is large enough so that we have $\angle(T_{X_1}M, \hat{T}_1) \leq 1/24$ on A_1. In particular, note that Item (i) is satisfied on A_1.

Let us now bound the probability that Item (ii) does not occur. As in Lemma A.10, we assume $\varepsilon_2 := \left(A \frac{f_{\max}^4 \log n}{f_{\min}^5 (n-1)} \right)^{\frac{2}{d+1}}$, where A is to be fixed later. For $x \in \mathrm{B}(X_1, r_0) \cap \partial M$, denote by $O_x^{int} = \pi_{\hat{T}_1}(x - X_1) - r_0 \eta_1^*(x)$.

Recall here that $\mathcal{Y}_1$ is defined by $\mathcal{Y}_1 := \pi_{T_1}(\mathrm{B}(X_1, R_0) \cap \mathcal{X}_n - X_1)$. If $\Omega \in \hat{T}_1$ is such that $\mathrm{B}(\Omega, \rho) \cap \mathcal{Y}_1 = \emptyset$ and $\|\Omega - O_x^{int}\| \leq \rho + r_0 - \varepsilon_2$ for some $\rho \geq \rho_-$ and $\rho_- + r_0 > \varepsilon_2 > 0$, then choosing $\Omega_0 = \Omega + (\rho - \rho_-)\frac{O_x^{int} - \Omega}{\rho + r_0 - \varepsilon_2}$ yields that

$$\begin{cases} \mathrm{B}(\Omega_0, \rho_-) \cap \mathcal{Y}_1 \subset \mathrm{B}(\Omega, \rho) \cap \mathcal{Y}_1 = \emptyset, \\ \|\Omega_0 - O_x^{int}\| \leq r_0 + \rho_- - \varepsilon_2. \end{cases}$$

But as $\|x - X_1\| \leq r_0$, Lemma A.7 ensures that on the event A_1 we have

$$\mathrm{B}(O_x^{in}, r_0) \cap \hat{T}_1 \subset \pi_{\hat{T}_1}(\mathrm{B}(X_1, 5r_0/2 + r_0) \cap M - X_1) \subset \pi_{\hat{T}_1}(\mathrm{B}(X_1, R_0) \cap M - X_1).$$

Thus, if we let

$$\mathcal{Q}_{r,\rho,\varepsilon} := \Big\{ (O, \Omega) \in \mathrm{B}_{\hat{T}_1}(0, 2r_0) \times \mathrm{B}_{\hat{T}_1}(0, 4r_0) \big| \|\Omega - O\| \leq r + \rho - \varepsilon \\ \text{and } \mathrm{B}_{\hat{T}_1}(O, r) \subset \pi_{\hat{T}_1}(\mathrm{B}(X_1, R_0) \cap M - X_1) \Big\},$$

then for all $\rho \geq \rho_-$, we have the inclusion of events

$$\begin{aligned}&\left\{\exists (x,\Omega) \in \mathrm{B}(X_1, r_0) \times \hat{T}_1 \mid \|\Omega - O_x^{int}\| \leq r_0 + \rho - \varepsilon_2 \text{ and } \mathrm{B}(\Omega,\rho) \cap \mathcal{Y}_1 = \emptyset\right\} \cap A_1 \\ &\quad \subset \bigcup_{(O,\Omega)\in \mathcal{Q}_{r_0,\rho_-,\varepsilon_2}} \{\mathrm{B}(\Omega,\rho_-) \cap \mathcal{Y}_1 = \emptyset\} \cap A_1.\end{aligned}$$

This union of events being infinite, we now discretize space by considering an $(\varepsilon_2/8)$-covering $\mathcal{C}(\varepsilon_2)$ of $\mathrm{B}_{\hat{T}_1}(0, 4r_0)$. For all $(\Omega, O) \in \mathcal{Q}_{r_0,\rho_-,\varepsilon_2}$, we also let Ω' and O' denote the closest elements in $\mathcal{C}(\varepsilon_2)$ to Ω and O respectively. Letting $r_0' := r_0 - \varepsilon_2/8$ and $\rho_0' := \rho_- - \varepsilon_2/8$, triangle inequality yields that on A_1,

$$\begin{cases} \mathrm{B}_{\hat{T}_1}(O', r_0') \subset \pi_{\hat{T}_1}(\mathrm{B}(X_1, R_0) \cap M) - X_1, \\ \mathrm{B}(\Omega', \rho_0') \cap \mathcal{Y}_1 = \emptyset, \\ \|\Omega' - O'\| \leq r_0' + \rho_0' - \varepsilon_2/2. \end{cases}$$

As a result, provided that n is large enough so that $\varepsilon_2 \leq 4r_0$, the previous event union satisfies

$$\bigcup_{\mathcal{Q}_{r_0,\rho_-,\varepsilon_2}} \{\mathrm{B}(\Omega,\rho_-) \cap \mathcal{Y}_1 = \emptyset\} \cap A_1 \subset \bigcup_{\mathcal{Q}_{\frac{r_0}{2},\frac{\rho_-}{2},\frac{\varepsilon_2}{2}} \cap \mathcal{C}(\varepsilon_2)^2} \left\{\mathrm{B}\left(\Omega, \frac{\rho_-}{2}\right) \cap \mathcal{Y}_1 = \emptyset\right\} \cap A_1.$$

Let $(O,\Omega) \in \mathcal{P}\left(\frac{r_0}{2}, \frac{\rho_-}{2}, \frac{\varepsilon_2}{2}\right)$ be now fixed. Recalling that $Y_i = X_i \mathbb{1}_{X_i \in \mathrm{B}(X_1,h)}$, and that the event A_1 is $\sigma(Y_2, \ldots, Y_n)$-measurable, we may write

$$\begin{aligned}&\mathbb{P}\left(A_1 \cap \left\{\mathrm{B}\left(\Omega, \frac{\rho_-}{2}\right) \cap \mathcal{Y}_1 = \emptyset\right\}\right) \\ &= \mathbb{E}\left[\mathbb{P}\left(A_1 \cap \left\{\mathrm{B}\left(\Omega, \frac{\rho_-}{2}\right) \cap \mathcal{Y}_1 = \emptyset\right\} \mid (Y_2, \ldots, Y_n)\right)\right] \\ &= \mathbb{E}\left[\mathbb{1}_{A_1}\mathbb{P}\left(\left\{\mathrm{B}\left(\Omega, \frac{\rho_-}{2}\right) \cap \mathcal{Y}_1 = \emptyset\right\} \mid (Y_2, \ldots, Y_n)\right)\right] \\ &\leq \mathbb{E}\left[\mathbb{1}_{A_1}\mathbb{P}\left(\min_{X_i \in \mathbb{X}_n \cap \mathrm{B}(X_1,R_0)} \left\|\pi_{\hat{T}_1}(X_i - X_1) - \Omega\right\| > \frac{\rho_-}{2} \mid (Y_2, \ldots, Y_n)\right)\right] \\ &\leq \mathbb{E}\left[\mathbb{1}_{A_1}\mathbb{P}\left(\min_{X_i \in \mathbb{X}_n \cap (\mathrm{B}(X_1,R_0) \setminus \mathrm{B}(X_1,h))} \left\|\pi_{\hat{T}_1}(X_i - X_1) - \Omega\right\| > \frac{\rho_-}{2} \mid (Y_2, \ldots, Y_n)\right)\right].\end{aligned}$$

Furthermore, as the family $(\pi_{\hat{T}_1}(X_i))_{X_i \notin \mathrm{B}(X_1,h)}$ is i.i.d conditionally on $(Y_2, \ldots, Y_n)$, Lemma A.10 yields

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{A_1}\mathbb{P}\left(\min_{X_i \in \mathbb{X}_n \cap (\mathrm{B}(X_1,R_0) \setminus \mathrm{B}(X_1,h))} \left\|\pi_{\hat{T}_1}(X_i - X_1) - \Omega\right\| > \frac{\rho_-}{2} \mid (Y_2, \ldots, Y_n)\right)\right] \\ &\leq \mathbb{E}\left[\mathbb{1}_{A_1}\left(1 - Ar_0^{\frac{d-1}{2}} C_d \frac{f_{\max}^4 \log n}{f_{\min}^4 (n-1)}\right)^{n - |\mathbb{X}_n \cap \mathrm{B}(X_1,h)|}\right] \\ &\leq \sum_{k=0}^{n-1} \binom{n-1}{k} (C_d f_{\max} h^d)^k \left(1 - Ar_0^{\frac{d-1}{2}} C_d \frac{f_{\max}^4 \log n}{f_{\min}^4 (n-1)}\right)^{n-k} \\ &\leq \left(1 - Ar_0^{\frac{d-1}{2}} C_d \frac{f_{\max}^4 \log n}{f_{\min}^4 (n-1)} + \frac{C_d f_{\max}^5 \log n}{f_{\min}^5 (n-1)}\right)^{n-1}.\end{aligned}$$

Choosing $A := C_d \frac{f_{\max}}{f_{\min}} r_0^{\frac{1-d}{2}} \geq C_d r_0^{\frac{1-d}{2}}$, yields that

$$\left(1 - Ar_0^{\frac{d-1}{2}} C_d \frac{f_{\max}^4 \log n}{f_{\min}^4 (n-1)} + \frac{C_d f_{\max}^5 \log n}{f_{\min}^5 (n-1)}\right)^{n-1} \leq n^{-C_d},$$

so that, for C_d large enough,

$$|\mathcal{C}(\varepsilon_2)|^2 \, \mathbb{P}\left(A_1 \cap \left\{\mathrm{B}\left(\Omega, \frac{\rho_-}{2}\right) \cap \mathcal{Y}_1 = \emptyset\right\}\right) \leq \left(\frac{1}{n}\right)^{1+\frac{2}{d}},$$

for n large enough. Thus, a union bound gives the result of Proposition 5.2, since we have set $\varepsilon_2 = r_0 \left(C_d \frac{f_{\max}^5}{f_{\min}^5} \frac{\log n}{f_{\min}(n-1) r_0^d}\right)^{\frac{2}{d+1}}$ for C_d large enough.

□

5.4 Proof of Theorem 3.15

The proof of Theorem 3.15 is based on the following deterministic result, whose proof is deferred to Appendix E.

Theorem 5.6 (Estimation with Local Linear Patches)**.** *Write $r_0 := (\tau_{\min} \wedge \tau_{\partial,\min})/40$, let $\varepsilon_0, a, \delta \geq 0$, and $0 \leq \theta, \theta' \leq 1/16$. Assume that we have:*

1. *A point cloud $\mathcal{X}_n \subset M$ such that $\mathrm{d_H}(M, \mathcal{X}_n) \leq \varepsilon_0$,*
2. *Estimated tangent spaces $(T_i)_{1 \leq i \leq n}$ such that $\max_{1 \leq i \leq n} \angle(T_{X_i} M, T_i) \leq \theta$,*
3. *A subset of boundary observations $\mathcal{X}_\partial \subset \mathcal{X}_n$ such that*

$$\max_{x \in \partial M} \mathrm{d}(x, \mathcal{X}_\partial) \leq \delta \text{ and } \max_{x \in \mathcal{X}_\partial} \mathrm{d}(x, \partial M) \leq a\delta^2,$$

from which we build interior observations

$$\mathring{\mathcal{X}}_{\varepsilon_{\partial M}} := \{X_i \in \mathcal{X}_n \mid \mathrm{d}(X_i, \mathcal{X}_\partial) \geq \varepsilon_{\partial M}/2\}.$$

4. *Estimated unit normal vectors $(\eta_i)_{1 \leq i \leq n}$ on $\mathcal{X}_\partial$ such that $\max_{X_i \in \mathcal{X}_\partial} \|\eta_i - \eta_{\pi_{\partial M}(X_i)}\| \leq \theta'$.*

Let $\mathbb{M} = \mathbb{M}(\mathcal{X}_n, \mathcal{X}_\partial, T, \eta)$ be defined as $\mathbb{M} := \mathbb{M}_{\mathrm{Int}} \cup \mathbb{M}_\partial$, with

$$\mathbb{M}_{\mathrm{Int}} := \bigcup_{X_i \in \mathring{\mathcal{X}}_{\varepsilon_{\partial M}}} X_i + \mathrm{B}_{T_i}(0, \varepsilon_{\mathring{M}}),$$
$$\mathbb{M}_\partial := \bigcup_{X_i \in \mathcal{X}_\partial} (X_i + \mathrm{B}_{T_i}(0, \varepsilon_{\partial M})) \cap \{z, \langle z - X_i, \eta_i \rangle \leq 0\},$$

Then if $\varepsilon_{\partial M} \leq r_0/2$, $\varepsilon_0 \leq \varepsilon_{\mathring{M}} \leq \varepsilon_{\partial M}/6$, and $\max\left\{\delta, a\delta^2\right\} \leq \varepsilon_{\partial M}/6$, we have

$$\mathrm{d_H}\big(M, \mathbb{M}\big) \leq \begin{cases} \varepsilon_{\mathring{M}} \left(\theta + \varepsilon_{\mathring{M}}/\tau_{\min}\right) & \text{if } \partial M = \emptyset, \\ 2a\delta^2 + 8\varepsilon_{\partial M} \left(\theta + \theta' + \varepsilon_{\partial M}/r_0\right) & \text{if } \partial M \neq \emptyset. \end{cases}$$

Equipped with Theorem 5.6, choose, for $i \in \{1, \ldots, n\}$, $T_i = \hat{T}_i$ as in Proposition 3.2, $\eta_i = \tilde{\eta}_i$ as in Theorem 3.7, and $\mathcal{X}_\partial = \mathcal{Y}_{R_0,r,\rho}$. Then we define

$$\hat{M} := \mathbb{M}(\mathbb{X}_n, \mathcal{Y}_{R_0,r,\rho}, \hat{T}, \tilde{\eta}).$$

Combining Proposition 3.2, Corollary 3.11, Theorem 3.7 and Lemma A.9 ensure that the requirements of Theorem 5.6 are satisfied with probability at least $1 - 4n^{-\frac{2}{d}}$ for n large enough, with the following choices of parameters: $\varepsilon_{\partial M} = 6\delta$,

$$\delta = 3r, \qquad \varepsilon_0 = \left(C_d \tfrac{\log n}{f_{\min} n}\right)^{\frac{1}{d}}, \qquad \varepsilon_{\mathring{M}} = \left(C_d \tfrac{\log n}{f_{\min} n}\right)^{\frac{1}{d}},$$
$$\theta = \left(C_d \tfrac{f_{\max}^{4+d}}{f_{\min}^{5+d}} \tfrac{\log n}{(n-1)\tau_{\min}^d}\right)^{\frac{1}{d}}, \quad \theta' = \tfrac{20r}{\sqrt{(\tau_{\min} \wedge \tau_{\partial,\min}) R_0}}, \quad a = (4(\tau_{\min} \wedge \tau_{\partial,\min}))^{-1},$$

which concludes the proof of the first bound in Theorem 3.15.

To get the bound in expectation, let K denote the diameter of M, and note that there exists $X_{i_0} \in \mathbb{X}_n$ such that $\{X_{i_0}\} \subset \hat{M}$, so that $\sup_{x \in M} \mathrm{d}(x, \hat{M}) \leq K$, almost surely. Conversely, since $\hat{M} \subset M + \mathrm{B}(0, \varepsilon_{\partial M} \vee \varepsilon_{\mathring{M}})$, we deduce that $\sup_{x \in \hat{M}} \mathrm{d}(x, M) \leq K$ for n large enough. Finally, noticing that for n large enough, the result follows by writing

$$\left(4n^{-\frac{2}{d}}\right) K \leq C_d(\tau_{\min} \wedge \tau_{\partial,\min}) \left[\left(\frac{f_{\max}^{2+d/2}}{f_{\min}^{2+d/2}} \frac{\log n}{f_{\min} \tau_{\min}^d n} \right)^{\frac{2}{d}} \wedge \left(\frac{f_{\max}^5}{f_{\min}^5} \frac{\log n}{f_{\min}(\tau_{\min} \wedge \tau_{\partial,\min})^d n} \right)^{\frac{2}{d+1}} \right].$$

6 Computational considerations and experimental illustrations

6.1 Pseudo-code and computational complexity

The algorithm `Boundary Structure` displays the pseudo-code for building the filtered set $\mathcal{Y}_{R_0,r,\rho}$ from Definition 3.3, alongside with the estimated unit normals from Definition 3.5. Note that it only uses standard algorithmic sub-blocks, such as PCA and Voronoi diagrams.

Its time complexity critically depends on the number of vertices K of the individual d-dimensional Voronoi cells. In worst case, each vertex corresponds to a d-simplex in the Delaunay triangulation, so that $K = O(n^{\lceil \frac{d}{2} \rceil})$, which corresponds to the maximum number of simplices in a Delaunay triangulation based on n points). Yet, for random point clouds, the typical number of such vertices reduces to $K = O\big(d^{\frac{d}{2}-1}\big)$ [41, Theorem 7.2, case $s = 0$].

At the end of the day, the average time complexity of `Boundary Structure` does not exceed (up to logarithmic factors)

$$O\left(Dn^2 + n\Big(N_{\mathrm{PCA}} Dd + N_{\mathrm{loc}} Dd + N_{\mathrm{loc}} 2^d K + (1 + N_{\mathrm{wit}}) dK\Big)\right),$$

where $N_{\mathrm{PCA}} = nh^d$, $N_{\mathrm{loc}} = nR_0^d$, and $N_{\mathrm{wit}} = nr^d$.

The parameters leading to optimal rates in our theoretical results (see Proposition 3.2 and Theorem 3.7) correspond to $h \asymp (1/n)^{1/d}$, $R_0 \asymp \rho \asymp 1$, and $r \asymp (1/n)^{1/(d+1)}$ up to $\log n$ factors. Hence, the above time complexity bound boils down to

$$O\left(n^2\big(Dd + d^{\frac{d}{2}}\big)\right).$$

Overall, note that the dependency on the ambient dimension D is limited to a linear factor. On the other hand, the leading factor in terms of sample size arises from the computation of the whole distance matrix of $\mathbb{X}_n$. This dependency could be mitigated to $O(DnN_{\mathrm{loc}})$ by only computing the local distances $(\|X_i - X_j\|)_{\|X_i - X_j\| \leq R_0}$. This can be done approximately, for instance via standard greedy exploratory geometric algorithms [32]. The factor $O(n^2)$ factor is also contributed from the fact that $N_{\mathrm{loc}} \asymp n$, since $R_0 \asymp 1$. Strategies consisting in taking $R_0 = o(1)$ could lead to a computation-precision tradeoff (see Theorem 3.12). Similarly, the choice $r = 0$ that we used in practice yield $N_{\mathrm{wit}} = 1$, further reducing algorithmic complexity.

Algorithm 1 `Boundary Structure` {*Average time complexity*}

Require:
$\mathbb{X}_n = \{X_1, \dots, X_n\} \subset \mathbb{R}^D$: Sample points
h: Bandwidth for principal component analysis {*write* $N_{\mathrm{PCA}} = nh^d$}
R_0: Macroscopic localization scale for projections {*write* $N_{\mathrm{loc}} = nR_0^d$}
r: Neighborhood radius of boundaryness witnesses {*write* $N_{\mathrm{wit}} = nr^d$}
ρ: Minimal width of witnessing Voronoi cells

Ensure: $h, r, R_0, \rho \geq 0$

Initialization with an empty boundary structure
$\mathcal{Y}_{R_0,r,\rho} \leftarrow \emptyset$
$\eta \leftarrow \emptyset$
Computation of pairwise point distances
$\mathrm{dist} \leftarrow \big(\|X_j - X_i\|\big)_{1 \leq i,j \leq n}$ {*time* $O(Dn^2)$}
for all $j \in \{1, \dots, n\}$ **do**
  Estimation of $T_{X_j}M$ via local PCA on a h neighborhood of X_j
  $\pi_{\hat{T}_j} \leftarrow \mathrm{PCA}_d(\mathbb{X}_n \cap \mathrm{B}(X_j, R_0))$ {*time* $O(N_{\mathrm{PCA}}Dd)$ *with fast-PCA [46]*}
  Projection of the R_0-neighborhood of X_j onto $\hat{T}_j$
  $\hat{\mathbb{X}}_n^{(j)} \leftarrow \pi_{\hat{T}_j}\big(\mathbb{X}_n \cap \mathrm{B}(X_j, R_0)\big)$ {*time* $O(N_{\mathrm{loc}}Dd)$}
  Computation of the Voronoi cells of the projected sample
  $\mathrm{Vor}_{R_0}^{(j)} \leftarrow \big(\mathrm{Vor}_{R_0}^{(j)}(X_i)\big)_{\hat{X}_i \in \hat{\mathbb{X}}_n^{(j)}}$ {*time* $O(N_{\mathrm{loc}}2^dK)$ *[47]*}
  for all $\hat{X}_i^{(j)} \in \hat{\mathbb{X}}_n^{(j)} \cap \mathrm{B}(X_j, r)$ **do**
    List the vertices $v_{i,1}, \dots, v_{i,K}$ of the Voronoi cell $\mathrm{Vor}_{R_0}^{(j)}(X_i)$ of $\hat{X}_i = \pi_{\hat{T}_j}(X_i)$
    {*Typical* $K = O\big(d^{\frac{d}{2}-1}\big)$ *[41, Theorem 7.2, case* $s = 0$*]*}
    Computation of the radius and prominent direction of $\mathrm{Vor}_{R_0}^{(j)}(X_i)$ when centered at $\pi_{\hat{T}_i}(X_j)$
    $\hat{v}_i^{(j)} \leftarrow \mathrm{argmax}_{1 \leq k \leq K} \|v_{i,k} - \pi_{\hat{T}_j}(X_i)\|$
    $\rho_i^{(j)} \leftarrow \|\hat{v}_i^{(j)} - \pi_{\hat{T}_j}(X_i)\|$ {*time* $O(dK)$}
    if $\rho_i^{(j)} > \rho$ **then**
      Addition of X_j to the boundary observations with associated unit normal
      $\mathcal{Y}_{R_0,r,\rho} \leftarrow \mathcal{Y}_{R_0,r,\rho} \cup \{X_j\}$
      $\eta_j \leftarrow \frac{\hat{v}_i^{(j)} - \pi_{\hat{T}_j}(X_i)}{\|\hat{v}_i^{(j)} - \pi_{\hat{T}_j}(X_i)\|}$
    end if
  end for {$O(1 + N_{\mathrm{wit}})$ *iterations*}
end for {n *iterations*}
return $(\mathcal{Y}_{R_0,r,\rho}, \eta)$

6.2 Heuristics for data-driven parameters calibration

6.2.1 Bandwidth h for PCA

Intuitively, the bandwidth h from Proposition 3.2 should be taken so that all the balls $\mathrm{B}(X_i, h)$ contain at least $d \vee \log n$ sample points. In practice, we set

$$h(k) = \inf\{r > 0 \mid \forall i \in \{1, \dots, n\}, |\mathrm{B}(X_i, r) \cap \mathbb{X}_n| \geq k\},$$

with $k = d \log n$. This particular scale is chosen so that, on average, the PCA's are computed using $\log n$ neighbors per estimated direction.

Let us note that in the noise-free case considered here, the choice of h (or k) does not impact significantly the tangent space estimation step provided $k \geq d+1$ and $h(k) = O(\mathrm{d_H}(M, \mathbb{X}_n))$. This latter Hausdorff distance can be approached by the smallest h such that $\mathrm{B}(\mathbb{X}_n, h)$ is connected. In the same spirit, pointwise choices of $h = h_i$ based on k-nearest neighbors could be a way to adapt to possible non-uniformity of sampling ($f_{\min} \ll f_{\max}$).

6.2.2 Macroscopic localization scale R_0 for projection

Throughout the theoretical analysis of the method, the scale R_0 of Definition 3.3 is chosen so that the (approximate) tangent projections $\pi_{\hat{T}_i} : M \cap \mathrm{B}(X_i, R_0) \to \hat{T}_i$ do not distort the metric significantly. Assessing this information empirically may be performed graphically, via the scatter-plot of all pairs

$$\left(\|X_i - X_j\|, \left\|\hat{T}_i(X_i - X_j))\right\|\right)_{1 \leq i,j \leq n}.$$

With this graphical representation in mind, a natural choice of R_0 is the largest radius R such that the plot remains close to the diagonal $y = x$ over $[0, R]$, where "close" needs to be properly defined.

This heuristic is used in the experiments of Section 6.3. As metric distortion is a multiplicative quantity, we consider

$$\gamma(R) := \min_{\|X_i - X_j\| \leq R} \frac{\left\|\hat{T}_i(X_i - X_j))\right\|}{\|X_i - X_j\|},$$

and we choose the largest R such that $\gamma(R)$ remains close to 1. That is, we pick

$$R_0 = \max\left\{ R > 0 \,\middle|\, \forall X_i, X_j \in \mathbb{X}_n \text{ s.t. } \|X_i - X_j\| \leq R, \left| \frac{\left\|\hat{T}_i(X_i - X_j))\right\|}{\|X_i - X_j\|} - 1 \right| \leq \delta \right\},$$

where $\delta < 1$ is a metric distortion tolerance parameter. We believe that any choice of $\delta \leq 1/2$ would have the method work.

6.2.3 Neighborhood radius r of boundaryness witnesses

We strongly believe that parameter r introduced in Definition 3.3 is a purely theoretical artifact. That is, the overall method is likely to have the same theoretical guarantees when applied with $r = 0$. We did not succeed in proving this conjecture, apart for $d = 1$, in which case ∂M consists of only two points per connected component of M.

On the practical side, we conducted our experiments with $r = 0$, which yielded satisfactory results. Recall that the smaller r, the smaller the set of detected boundary points $\mathcal{Y}_{R_0,r,\rho}$, which could lead to potentially too many false negative (i.e. nearby-boundary points missed).

Yet, beyond the present idealized framework where points are not corrupted with noise, this parameter may have an influence. If so, a possible calibration strategy could consist in investigating the size of the Voronoi cell of $\pi_{\hat{T}_j}(X_i - X_j)$, where X_j ranges among the k-nearest neighbors of X_i for growing k's, and to stop when the number of detected points stabilizes.

6.2.4 Minimal width ρ of witnessing Voronoi cells

We finally move to discussion about ρ, the last parameter involved in Definition 3.3. Recall that for $r = 0$, the labelling of X_i is only based on the size of the Voronoi cell of $\pi_{\hat{T}_i}(X_i)$. Hence, for a given point X_i, we are led to compute

$$\rho_i = \max_{1 \leq k \leq K} \|\pi_{\hat{T}_i}(X_i) - v_{i,k}\|,$$

where $(v_{i,k})_{k \leq K}$ are the vertices of the Voronoi cell (see `Boundary Structure`). For interior points, ρ_i is expected to be small, while for points close to the boundary, ρ_i is expected to be larger. The value of ρ effectively fixes the chosen cutoff between "small" and 'large" cells. To pick ρ wisely, we investigate the distribution of the values $(\rho_i)_{1 \leq i \leq n}$.

That is, we reorder values $\rho_{(1)}, \ldots, \rho_{(n)}$, and we plot the graph $(i, \rho_{(i)})_{1 \leq i \leq n}$. This graph typically exhibits a sharp jump (see Tables 1 and 2). As the value of this jump corresponds to a phase transition between the two regimes we want to distinguish, we select ρ to be the mid-value of this first jump. Note that for $r > 0$, a similar strategy based on the $\rho_i^{(j)}$'s (as defined in Section 6.1) can easily be built.

6.3 Simulations on low-dimensional examples

We now illustrate the boundary detection, normal vector estimation and parameter tuning heuristics on some examples on four low-dimensional examples. Namely, the toy distributions that we consider consist of the uniform distributions over the following sets.

($d = 1$, $D = 3$) The spiral given by the parametrization $[0, 5\pi] \ni \theta \mapsto (\cos(\theta), \sin(\theta), \theta/3)$.

($d = 2$, $D = 2$) The annulus $\mathrm{B}(0, 1) \setminus \mathrm{B}(0, 0.4)$.

($d = 2$, $D = 3$) The unit half sphere $\{(x, y, z) |\ x^2 + y^2 + z^2 = 1 \text{ and } x \geq 0\}$.

($d = 2$, $D = 3$) The Möbius strip given by the parametrization

$$[-1, 1] \times [0, 2\pi] \ni (u, \theta) \mapsto ((u \cos(\theta/2) + 3) \cos(\theta), (u \cos(\theta/2) + 3) \sin(\theta), u \sin(\theta/2)).$$

For each of these four distributions, Tables 1 and 2 present:

- The metric distortion scatterplot used for the calibration of R_0 (see Section 6.2.2);
- The order distance histogram plot used for the calibration of ρ (see Section 6.2.4);
- Points clouds of $\{500, 1000, 2000, 5000\} \ni n$-samples, with the detected boundary observations and their associated estimated normals displayed in red.

In all the cases, bandwidth h is chosen using the rule of Section 6.2.1, and $r = 0$ (see Section 6.2.3). These plots are for illustrative purpose only, and are not meant to illustrate minimax convergence rates. Qualitatively, let us point the following:

- For the spiral, exactly two observations are labelled as boundary observations with associated Voronoi cells that are unbounded. This advocates that for possibly setting $\rho = \infty$ in the one-dimensional case $d = 1$.

- For the annulus, no tangent projection is performed, since we are on a full-dimensional domain ($d = D$). This is why the scatterplot of $\left(\|X_i - X_j\|, \left\|\hat{T}_i(X_i - X_j))\right\|\right)_{1 \leq i,j \leq n}$ coincides with the identity. This advocates for setting $R_0 = \infty$ in the full-dimensional case $d = D$. Note that if so, only one global Voronoi diagram (that of the complete sample points $\mathbb{X}_n$) needs to be computed, as opposed to one local Voronoi per point.

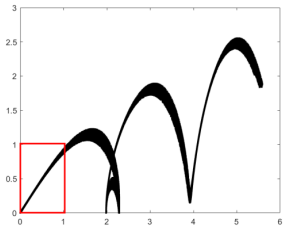
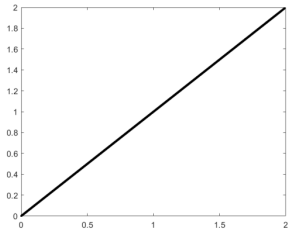
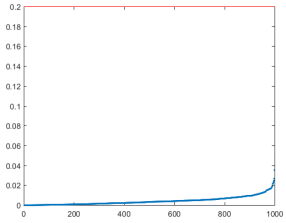
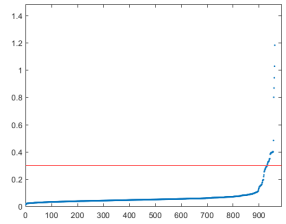
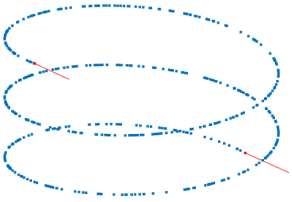
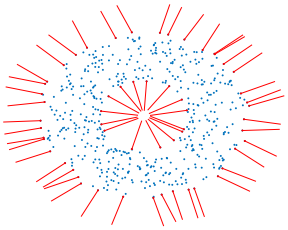
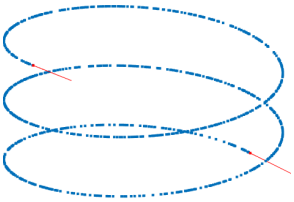
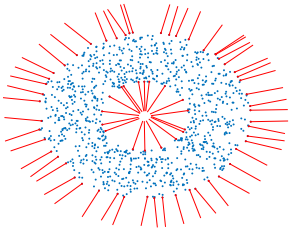
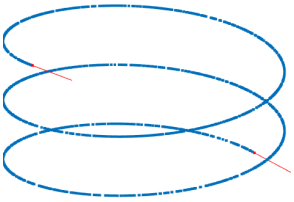
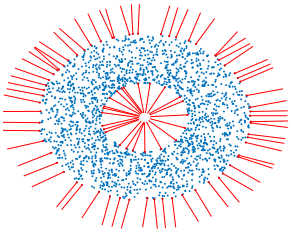
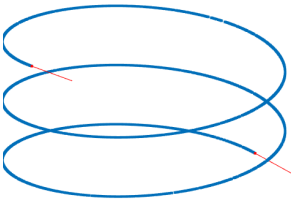
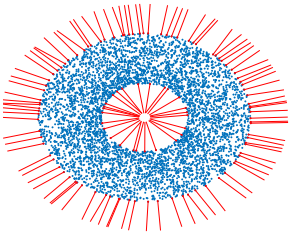
M	Spiral	Annulus
Calibration of R_0		
Calibration of ρ		
$n = 500$		
$n = 1000$		
$n = 2000$		
$n = 5000$		

Table 1: Simulations results for the spiral and the annulus.

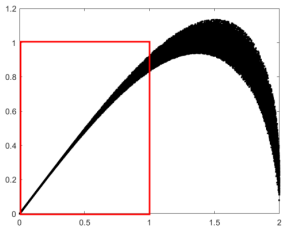
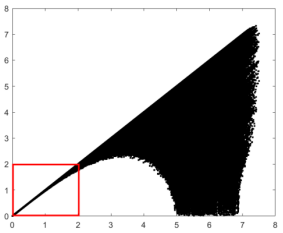
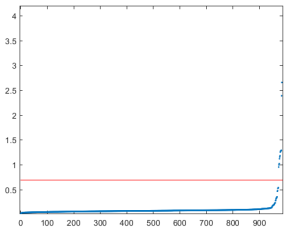
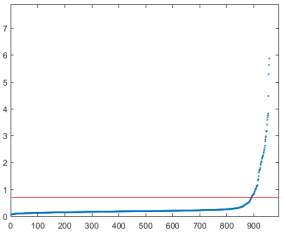
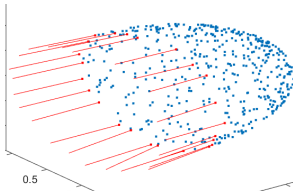
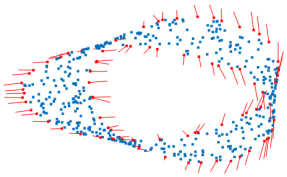
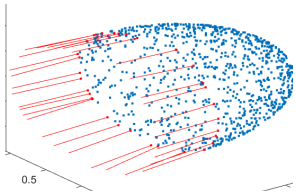
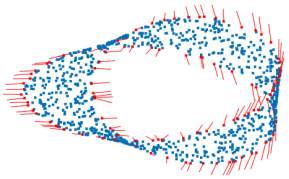
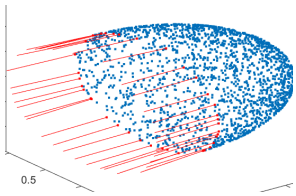
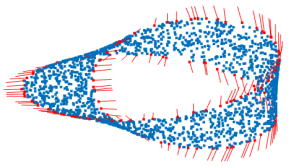
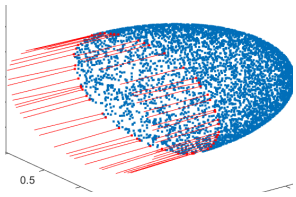
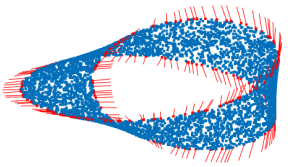
M	Half-sphere	Möbius strip
Calibration of R_0		
Calibration of ρ		
$n = 500$		
$n = 1000$		
$n = 2000$		
$n = 5000$		

Table 2: Simulations results for the half-sphere and the Möbius strip.

A A shortlist of intermediate geometric results

This section gathers the main geometric results that are of use in the main derivations (Section 5). For the sake of concision, proofs of these results are given in Appendix B. Throughout, $\mathbb{G}^{D,d}$ stands for the Grassmannian — i.e. the space of d-dimensional linear subspaces of $\mathbb{R}^D$ —, and d_S for the geodesic distance of $S \subset \mathbb{R}^D$.

A.1 Geodesics and tangent spaces

We begin with a result that connects geodesic and Euclidean distance.

Lemma A.1 (Geodesic Bounds)**.** *Let $p, q \in M$ such that $\|p - q\| \leq \tau_{\min}$. Then*

$$\|p - q\| \leq \mathrm{d}_M(p, q) \leq 2\|p - q\|.$$

A short proof is given in Appendix B.1. This result is well-known in the empty boundary case (see [2, Proposition 8.6]). In the general case, Lemma A.1 follows from [13, Lemma 3]. The last result of this section connects tangent spaces variations with the geodesic distance between their base points.

Proposition A.2 (Tangent Space Stability)**.** *Let $M \in \mathcal{M}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}$. Then, for $x, y \in M$,*

$$\angle(T_xM, T_yM) \leq \mathrm{d}_M(x, y)/\tau_M.$$

If $\partial M \neq \emptyset$, then for all $p, q \in \partial M$,

$$\angle(T_p\partial M, T_q\partial M) \leq \mathrm{d}_{\partial M}(p, q)/\tau_{\partial M}.$$

A proof of Proposition A.2 is given in Appendix B.1. Combining the two angle bounds from Proposition A.2 easily yields a bound on the angle between the *linear spaces* $\mathrm{span}(\eta_p)$ and $\mathrm{span}(\eta_q)$, for $p, q \in \partial M$. Actually, making use of the structure of normal *cones*, a bound on $\|\eta_p - \eta_q\|$ can be derived, as presented below.

Proposition A.3 (Normal Vector Stability)**.** *Let $M \in \mathcal{M}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}$. Then for all $p, q \in \partial M$ such that $\|p - q\| \leq (\tau_M \wedge \tau_{\partial M})/32$, we have*

$$\|\eta_p - \eta_q\| \leq 9\|p - q\|/(\tau_M \wedge \tau_{\partial M}).$$

A proof of Proposition A.3 may be found in Appendix B.1.

A.2 Projections

Projections onto tangent spaces and normal directions play a key role in the estimation schemes on this work. First, we adapt [27, Theorem 4.18] to the case where a small perturbation of the tangent space is allowed.

Proposition A.4 (Tangent and Normal Components of Increments)**.** *Let $x, y \in M$, and $T \in \mathbb{G}^{D,d}$ be such that $\angle(T_xM, T) \leq \theta$. Write $(x - y)^T$ and $(x - y)^\perp$ for the orthogonal projection of $x - y$ onto T and $T^\perp$ respectively. Then,*

$$\begin{aligned}
\|(y - x)^\perp\| &\leq \|y - x\| \left(\theta + \|y - x\| / (2\tau_{\min})\right), \\
\|(y - x)^T\| &\geq \|y - x\| \left(1 - \theta - \|y - x\| / (2\tau_{\min})\right).
\end{aligned}$$

A proof of Proposition A.4 is given in Appendix B.2. The following result ensures that estimates of the normal direction to the boundary may be derived from a suitable tangent space estimator.

Proposition A.5 (Normals from Tangent Spaces)**.** *Let $x \in \partial M$, and $T \in \mathbb{G}^{D,d}$ such that $\angle(T_x M, T) < 1$. Then $T \cap Nor(x, M)$ contains a unique unit vector η, and it satisfies*

$$\|\eta - \eta_x\| \leq \sqrt{2}\angle(T_x M, T).$$

A proof of Proposition A.5 can be found in Appendix B.2. The remaining results of this section describe the structure of the projection of balls onto perturbed tangent spaces. We begin by investigating the case where the center of the ball is not on the boundary.

Lemma A.6 (Far-Boundary Balls)**.** *Let $x \in M$ and $T \in \mathbb{G}^{D,d}$ be such that $\angle(T_x M, T) \leq \theta \leq 1/8$. If $\mathrm{d}(x, \partial M) > 0$ (with the convention $\mathrm{d}(x, \emptyset) = +\infty$), and $R \leq \tau_{\min}/16$, then*

$$\mathrm{B}_T\left(0, \frac{4}{5}\min\left\{R, \mathrm{d}(x, \partial M)\right\}\right) \subset \pi_T(\mathrm{B}(x, R) \cap M - x).$$

A proof of Lemma A.6 is given in Appendix B.3. Next, Lemma A.7 describes $\pi_T(\mathrm{B}(x, R) \cap M - x)$ whenever x is a boundary point.

Lemma A.7 (Near-Boundary Balls)**.** *Assume that $\partial M \neq \emptyset$. Let $x \in \partial M$ and $T \in \mathbb{G}^{D,d}$ be such that $\angle(T_x M, T) \leq \theta \leq 1/8$. Denote by $\hat{\eta}$ the unit vector of $T \cap Nor(x, M)$, choose $R \leq \tau_{\min}/16$ and $r \leq \min\{2R/5, 7\tau_{\partial,\min}/5\}$.*

Then, writing $O^{in} := -r\hat{\eta}$ and $O^{out} := r\hat{\eta}$, we have

$$\mathrm{B}(O^{in}, r) \cap T \subset \pi_T(\mathrm{B}(x, R) \cap M - x) \subset \mathring{\mathrm{B}}(O^{out}, r)^c \cap T.$$

A proof of Lemma A.7 may be found in Appendix B.3. A consequence of Lemma A.7 is the following Corollary A.8, that will be useful in the proof of Theorem 5.1.

Corollary A.8 (Parallelism of Projected Normals)**.** *Assume that $\partial M \neq \emptyset$. Let $x \in M$ be such that $\mathrm{d}(x, \partial M) < \tau_{\min}/16$, and $y \in \mathbb{R}^D$. For $T \in \mathbb{G}^{D,d}$, let $x^* \in \pi_{y+T}(\partial M \cap \mathrm{B}(x, \tau_{\min}/16))$ be any point such that*

$$\|x^* - \pi_{y+T}(x)\| = \mathrm{d}(\pi_{y+T}(x), \pi_{y+T}(\partial M \cap \mathrm{B}(x, \tau_{\min}/16)),$$

and

$$x' \in \partial M \cap \mathrm{B}(x, \tau_{\min}/16) \text{ such that } \pi_{y+T}(x') = x^*.$$

If $\angle(T_{x'} M, T) \leq 1/8$, then $Nor(x', M) \cap T$ contains a unique unit vector $\eta^*(x')$, and

$$x^* - \pi_{y+T}(x) = \|x^* - \pi_{y+T}(x)\|\, \eta^*(x').$$

A proof of Corollary A.8 is given in Appendix B.3.

A.3 Covering and volume bounds

This last preliminary section provides probabilistic bounds on the sampling density of $\mathbb{X}_n$ in M, and bounds on the volume of intersection of balls. They will drive the convergence rates of Theorem 3.7. First, we adapt [2, Lemma 9.1] to the non empty boundary case.

Lemma A.9 (Sampling Density Bound)**.** *Let $\varepsilon_1 = \left(C_d \frac{\log n}{f_{\min} n}\right)^{\frac{1}{d}}$, for C_d large enough. Then, for n large enough so that $\varepsilon_1 \leq \frac{\tau_{\min}}{16} \wedge \frac{\tau_{\partial,\min}}{2}$, we have, with probability larger than $1 - n^{-3}$,*

$$\mathrm{d_H}(M, \mathbb{X}_n) \leq \varepsilon_1.$$

A proof of Lemma A.9 is given in Appendix B.4. It guarantees that the convergence rate of the sample $\mathbb{X}_n$, seen as a Hausdorff estimator of M, is the same as in the empty boundary case. Next, Lemma A.10 below provides bounds on the mass of projected intersection of balls.

Lemma A.10 (Mass of Intersection of Curved Balls)**.** *Let $x \in M$, and $T \in \mathbb{G}^{D,d}$. Let $O \in T$, and $r, R \geq 0$ be such that $\mathrm{B}_T(O,r) \subset \pi_T(\mathrm{B}(x,R) \cap M - x)$. For $A \geq C'_d r^{\frac{1-d}{2}}$, write*

$$h = \left(\frac{C_d f_{max}^4 \log n}{f_{\min}^5 (n-1)}\right)^{\frac{1}{d}}, \text{ and } \varepsilon_2 = \left(A \frac{f_{max}^4 \log n}{f_{\min}^5 (n-1)}\right)^{\frac{2}{d+1}}.$$

Then for n large enough, for all $\rho \geq r$ and $\Omega \in T$ such that $\|\Omega - O\| \leq r + \rho - \varepsilon_2$,

$$\int_{M \cap (\mathrm{B}(x,R) \setminus \mathrm{B}(x,h))} \mathbb{1}_{\pi_T(u-x) \in \mathrm{B}(O,r) \cap \mathrm{B}(\Omega,\rho)} f(u) \mathcal{H}^d(\mathrm{d}u) \geq A r^{\frac{d-1}{2}} C''_d \frac{f_{max}^4 \log n}{f_{\min}^4 (n-1)}.$$

A proof of Lemma A.10 can be found in Appendix B.4. From a sampling point of view, it will ensure that such intersections of (projected) balls will contain at least one sample point with high probability. This point will allow to detect and characterize the boundary observations (see Theorem 5.1).

B Geometric properties of manifolds with boundary

This Section gathers the proofs for Appendix A. To ease readability, statements are recalled before their proofs.

B.1 Geodesics and tangent space variations

In addition to the Euclidean structure induced by $\mathbb{R}^D$ on $M \subset \mathbb{R}^D$, we can also endow M and ∂M with their intrinsic geodesic distances d_M and $\mathrm{d}_{\partial M}$ respectively. To cover both cases at once, let $S \in \{M, \partial M\}$. Given a $\mathcal{C}^1$ curve $c : [a,b] \to S$, the length of c is defined as $\mathrm{Length}(c) = \int_a^b \|c'(t)\| \, dt$. Given $p, q \in S$ belonging to the same connected component of S, there always exists a path $\gamma_{p \to q}$ of minimal length joining p and q [16, Proposition 2.5.19]. Such a curve $\gamma_{p \to q}$ is called geodesic, and the geodesic distance between p and q is given by $\mathrm{d}_S(p,q) = \mathrm{Length}(\gamma_{p \to q})$. If x and y stand in different connected components of S, then $\mathrm{d}_S(x,y) = \infty$.

A geodesic γ such that $\|\gamma'(t)\| = 1$ for all t is called *arc-length parametrized*. Unless stated otherwise, we always assume that geodesics are parametrized by arc-length. If S has empty boundary, then for all $p \in S$ and all unit vectors $v \in T_p S$, we denote by $\gamma_{p,v}$ the unique arc-length parametrized geodesic of S such that $\gamma_{p,v}(0) = p$ and $\gamma'_{p,v}(0) = v$ [25, Chap. 7, Theorem 2.8]. The exponential map is then defined as $\exp_p^S(vt) = \gamma_{p,v}(t)$. Note that if in addition S is compact, $\exp_p^S : T_p S \to S$ is defined globally on $T_p S$ [16, Theorem 2.5.28]. We let $\mathrm{B}_S(p,s)$ denote the closed geodesic ball of center $p \in S$ and of radius $s \geq 0$.

Although they might differ drastically at long range, geodesic and Euclidean distances are good approximations of one another when evaluated between close enough points. The following result quantifies this intuition, and implies Lemma A.1.

Proposition B.1. *Let $S \subset \mathbb{R}^D$ have positive reach $\tau_S > 0$, and $x, y \in S$ be such that $\|y - x\| \leq \tau_S$. Then,*

$$\|y - x\| \leq \mathrm{d}_S(x,y) \leq \left(1 + \frac{\|y-x\|^2}{20 \tau_S^2}\right) \|y - x\|.$$

Proof of Proposition B.1. We clearly have $\|y-x\| \leq \mathrm{d}_S(x,y)$, and on the other hand, [13, Lemma 3] yields

$$\mathrm{d}_S(x,y) \leq 2\tau_S \arcsin\left(\frac{\|y-x\|}{2\tau_S}\right) \leq \left(1+\frac{\|y-x\|^2}{20\tau_S^2}\right)\|y-x\|,$$

where the last inequality follows uses that $\arcsin t \leq t(1+t^2/5)$ for all $0 \leq t \leq 1/2$. □

Next, we ensure that the angle between tangent spaces can be bounded in terms of geodesic distances between base points. In the empty boundary case, this result is well known, and can be shown using via parallel transportation of tangent vectors (see the proof of [3, Lemma A.1]). In the general case, the tangent space stability property writes as follows.

Proposition A.2 (Tangent Space Stability)**.** *Let $M \in \mathcal{M}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}$. Then, for $x, y \in M$,*

$$\angle(T_xM, T_yM) \leq \mathrm{d}_M(x,y)/\tau_M.$$

If $\partial M \neq \emptyset$, then for all $p, q \in \partial M$,

$$\angle(T_p\partial M, T_q\partial M) \leq \mathrm{d}_{\partial M}(p,q)/\tau_{\partial M}.$$

Proof of Proposition A.2. If $\partial M = \emptyset$, the first claim follows from [13, Lemma 6].

Assume that $\partial M \neq \emptyset$. From Proposition 2.3, ∂M is a $\mathcal{C}^2$-submanifold without boundary. Then, the second statement also directly follows from [13, Lemma 6]. For the first claim, the key technical point is to handle geodesics that would hit the boundary.

To do this we define a push-inwards operator that will allow to consider path in the interior of M only. First, an elementary results on an atlas of M is needed.

Lemma B.2. *Let $U_1, \ldots, U_k$ be charts of M that cover ∂M. Then there exists $r_0 > 0$ such that*

$$\forall p \in \partial M \quad \exists j \in \{1,\ldots,k\} \quad \mathring{\mathrm{B}}(p, r_0) \cap M \subset U_j \cap M.$$

We now consider a smooth kernel $K : \mathbb{R}_+ \to [0,1]$ such that

$$K(x) = \begin{cases} 1 & \text{if } x \leq \tau_{\partial M}/4 \\ 0 & \text{if } x \geq \tau_{\partial M}/2 \end{cases}$$

and we define the vector field $\mathbf{V}$ on M by

$$\mathbf{V}(p) := \begin{cases} K\left[\mathrm{d}(p,\partial M)\right] \pi_{T_pM}\left(\nabla_p\left(\mathrm{d}(\cdot,\partial M)\right)\right) & \text{if } \mathrm{d}(p,\partial M) < (r_0 \wedge \tau_{\partial M})/2, \\ 0 & \text{otherwise.} \end{cases}$$

Note that if $q \in \partial M$, $\nabla_q \mathrm{d}(\cdot,\partial M) = -\eta_q$, where η_q is the unit outward-pointing normal vector at q. By construction, $\mathbf{V}$ is a $\mathcal{C}^1$ tangent vector field on M. We now examine its flow.

Lemma B.3. *For all $p \in M$, the flow of $\mathbf{V}$ starting from p is defined globally on $\mathbb{R}_+$.*

Equipped with Lemma B.3, we may define our *push-inwards* operator as follows:

$$\begin{aligned} g_\varepsilon \colon M &\to M \\ p &\mapsto g(p,\varepsilon) \end{aligned}$$

where $g(p,t)$ denotes the flow of $\mathbf{V}$ at time $t \geq 0$ starting from $p \in M$. The following properties of g_ε will shortly be of technical interest.

Lemma B.4. *For all $p \in M$ and $\varepsilon > 0$,*

$$\|g_\varepsilon(p) - p\| \leq \varepsilon, \qquad g_\varepsilon(p) \notin \partial M, \qquad \text{and } \|d_p g_\varepsilon - Id_{T_pM}\|_{op} \leq K\varepsilon e^{K\varepsilon},$$

where $K = \sup_{p\in M} \|d_p \mathbf{V}\|_{op}$.

We can now finish the proof of the first result in Proposition A.2. We let $p, q \in M$, and γ a unit-speed curve joining p and q whith length $\mathrm{d}_M(p,q)$. We define γ_ε as the push-inwards of γ, that is

$$\gamma_\varepsilon(t) := g_\varepsilon(\gamma(t)),$$

for all $t \in [0, \mathrm{d}_M(p,q)]$. As $g_\varepsilon(p) \notin \partial M$ for all $p \in M$ (Lemma B.4), parallel transportation of tangent vectors in the interior $\operatorname{Int} M$ of M (see for instance the proof of [3, Lemma A.1]) yields that

$$\angle(T_{p_\varepsilon}M, T_{q_\varepsilon}M) \leq \frac{L(\gamma_\varepsilon)}{\tau_M},$$

where $p_\varepsilon = g_\varepsilon(p)$, $q_\varepsilon = g_\varepsilon(q)$, and $L(\gamma_\varepsilon)$ denotes the length of γ_ε. But from Lemma B.4 again,

$$L(\gamma_\varepsilon) = \int_0^{\mathrm{d}_M(p,q)} \|\gamma_\varepsilon'(t)\| \mathrm{d}t = \int_0^{\mathrm{d}_M(p,q)} \|d_{\gamma(t)} g_\varepsilon \left[\gamma'(t)\right]\| \mathrm{d}t \leq (1 + K\varepsilon e^{K\varepsilon}) \mathrm{d}_M(p,q).$$

$\angle(T_{p_\varepsilon M}, T_pM) \leq K\varepsilon e^{K\varepsilon}$, and $\angle(T_{q_\varepsilon}M, T_qM) \leq K\varepsilon e^{K\varepsilon}$. As a result, triangle inequality yields

$$\angle(T_pM, T_qM) \leq 2K\varepsilon e^{K\varepsilon} + (1 + K\varepsilon e^{K\varepsilon}) \frac{\mathrm{d}_M(p,q)}{\tau_M},$$

so that the result follows after letting $\varepsilon \to 0$. □

We finally prove the intermediate results of Lemmas B.2 to B.4 that we just used to derive Proposition A.2.

Proof of Lemma B.2. For all $p \in \partial M$, set

$$r(p) := \sup\{r \geq 0 \mid \exists j \in \{1, \dots, k\}, \mathring{\mathrm{B}}(p,r) \subset U_j\}.$$

Note that since $(U_i)_{1\leq i\leq k}$ is an open covering of ∂M we have $r(p) > 0$. Consider

$$r_0 := \inf_{p\in\partial M} r(p),$$

which clearly satisfies the announced statement by definition. Suppose, for contradiction, that $r_0 = 0$. Then there would exist a sequence $(p_n)_{n\in\mathbb{N}} \in (\partial M)^{\mathbb{N}}$ such that $r(p_n) \to 0$. As ∂M is compact, we may assume (up to extraction) that $p_n \to p \in \partial M$ as $n \to +\infty$. As a result, for n large enough, we have $\mathring{\mathrm{B}}(p_n, r(p_n)) \subset \mathring{\mathrm{B}}(p, r(p)) \subset U_{j_0}$ for some j_0, which is a contradiction. □

Proof of Lemma B.3. We distinguish cases according to the value of $\mathrm{d}(p, \partial M)$ with respect to the chart radius r_0 of Lemma B.2.

- If $\mathrm{d}(p, \partial M) \geq \tau_{\partial M}/2$, then $\mathbf{V}(p) = 0$ and the flow of $\mathbf{V}$ starting from p is $p(t) = p$ for all $t \geq 0$.
- If $r_0/2 < \mathrm{d}(p, \partial M) \leq \tau_{\partial M}/2$, then we may find $r_1 \in (0, r_0/2)$ such that $\mathring{\mathrm{B}}(p, r_1) \cap M$ is diffeomorphic to an open subset of $\mathbb{R}^d$. Using Cauchy-Lipschitz theorem in this chart space, we get that there exists $t_0 > 0$ such that the flow of $\mathbf{V}$ starting from p is well-defined at least on $[0, t_0)$.

- If $\mathrm{d}(p, \partial M) \leq r_0/2$, denote by $q = \pi_{\partial M}(p)$ and let $j \in \{1, \dots, k\}$ be such that $\mathring{\mathrm{B}}(q, r_0) \subset U_j$, where $\psi_j : U_j \cap M \to (\mathbb{R}^{d-1} \times \mathbb{R}_+) \cap \psi_j(U_j)$ is a chart of M. Without loss of generality we may assume that $d_q(\psi_j)(\eta_q) = -e_d$, where e_d is the d-th vector of the canonical basis of $\mathbb{R}^d$.

  Let $r_1 > 0$ be such that $V_1 = \mathring{\mathrm{B}}(\psi_j(p), r_1) \cap (\mathbb{R}^{d-1} \times \mathbb{R}_+) \subset (\mathbb{R}^{d-1} \times \mathbb{R}_+) \cap \psi_j(U_j)$, and denote by $\mathbf{V}_2$ the vector field on V_1 defined by $d\psi_j[\mathbf{V}]$. Then $\mathbf{V}_2$ can be extended into a Lipschitz vector field $\mathbf{V}_3$ on $\mathring{\mathrm{B}}(\psi_j(p), r_1)$, by choosing $\mathbf{V}_3(x_1, \dots, x_d) = \mathbf{V}_3(x_1, \dots, 0)$ if $x_d \leq 0$.

  Then, the Cauchy-Lipschitz theorem ensures that there exists t_0 such that the flow of $\mathbf{V}_3$ starting from $\psi_j(p)$ is defined on $]-t_0, t_0[$. Let $g_2(t, \psi_j(p))$ denote this flow. According to Lemma B.2, it holds $\langle \mathbf{V}_3(g_2(0, \psi_j(p))), e_d \rangle = 1$. Thus, there exists $t_1 > 0$ such that for all $t \in [0, t_1]$, $g_2(t, p) \in \mathring{\mathrm{B}}(\psi_j(p), r_1) \cap (\mathbb{R}^{d-1} \times \mathbb{R}_+)$, and therefore the flow of $\mathbf{V}_3$ starting from $\psi_j(p)$ stays in $\mathring{\mathrm{B}}(\psi_j(p), r_1) \cap (\mathbb{R}^{d-1} \times \mathbb{R}_+)$. When pushed back, this means that the flow of $\mathbf{V}$ starting from p stays in the chart (U_j, ψ_j).

In summary, we have shown that for all $p \in M$ there exists $t_p > 0$ such that the flow $g(t, p)$ of $\mathbf{V}$ starting from p is well-defined for $t \in [0, t_p]$. Since $g(\cdot, p)$ goes to the compact M and satisfies $g(t_1 + t_2, p) = g(t_2, g(t_1, p))$, we deduce that for all $p \in M$, $g(\cdot, p)$ is well-defined on $\mathbb{R}_+$. □

Proof of Lemma B.4. Since $\|\mathbf{V}\| \leq 1$, we directly get that

$$\|g_\varepsilon(p) - p\| = \left\| \int_0^\varepsilon \mathbf{V}(g(p,t)) \mathrm{d}t \right\| \leq \int_0^\varepsilon \|\mathbf{V}(g(p,t))\| \, \mathrm{d}t \leq \varepsilon.$$

To obtain the second point, write $\mathrm{d}(g_\varepsilon(p), \partial M) - \mathrm{d}(p, \partial M)$ as

$$\begin{aligned} &\int_0^\varepsilon \left\langle \mathbf{V}(g(p,t)), \nabla_{g(p,t)} \mathrm{d}(\cdot, \partial M) \right\rangle \mathrm{d}t \\ &\quad = \mathrm{d}(p, \partial M) + \int_0^\varepsilon K\left[\mathrm{d}(g(p,t), \partial M)\right] \left\langle \pi_{T_{g(p,t)}M}(\nabla_{g(p,t)} \mathrm{d}(\cdot, \partial M)), \nabla_{g(p,t)} \mathrm{d}(\cdot, \partial M) \right\rangle \mathrm{d}t. \end{aligned}$$

Thus,

- If $p \notin \partial M$, then $\mathrm{d}(g_\varepsilon(p), \partial M) \geq \mathrm{d}(p, \partial M) > 0$.
- If $p \in \partial M$, then $\pi_{T_pM}\left(\nabla_{g(p,0)} \mathrm{d}(\cdot, \partial M)\right) = -\eta_p$. Since $\mathbf{V}$ is continuous, there exists t_0 such that for all $t \leq t_0$, we have

$$\left\langle \mathbf{V}(g(p,t)), \nabla_{g(p,t)} \mathrm{d}(\cdot, \partial M) \right\rangle \geq 1/2 > 0.$$

As a result, we also get that $\mathrm{d}(g_\varepsilon(p), \partial M) > 0$ for all $\varepsilon > 0$.

For the third point, we write $K := \sup_{p \in M} \|d_p \mathbf{V}\|_{op} < \infty$, since $\mathbf{V}$ is $\mathcal{C}^1$ and M compact. Let $v \in T_pM$ be a unit vector, and γ be a path such that $\gamma(0) = p$ and $\gamma'(0) = v$. For a fixed t and $u \leq \varepsilon$, consider $f(u) := \|g(\gamma(t), u) - g(p, u)\|^2$. Then

$$\begin{aligned} |f'(u)| &= 2\left|\langle g(\gamma(t), u) - g(p, u), \mathbf{V}(g(\gamma(t), u)) - \mathbf{V}(g(p, u)) \rangle\right| \\ &\leq 2K f(u). \end{aligned}$$

Since $f(0) = \|\gamma(t) - p\|^2$, we deduce that $f(u) \leq \|\gamma(t) - p\|^2 e^{2Ku}$, so that

$$\|g(\gamma(t), u) - g(p, u)\| \leq \|\gamma(t) - p\| e^{Ku}.$$

But since

$$g_\varepsilon(\gamma(t)) - \gamma(t) = \int_0^\varepsilon \mathbf{V}(g(\gamma(t), u))\mathrm{d}u,$$

we have

$$g_\varepsilon(\gamma(t)) - g_\varepsilon(p) = tv + o(t) + \int_0^\varepsilon \left(\mathbf{V}(g(\gamma(t), u)) - \mathbf{V}(g(p, u))\right) \mathrm{d}u.$$

Thus

$$\left\| \frac{g_\varepsilon(\gamma(t)) - g_\varepsilon(p)}{t} - v \right\| \leq o(1) + K\varepsilon e^{K\varepsilon} \|\gamma(t) - p\|/t.$$

Letting $t \to 0$, we get that $\|d_p g_\varepsilon - Id_{T_pM}\| \leq K\varepsilon e^{K\varepsilon}$, since $\|\gamma(t) - p\|/t \to \|v\| = 1$. □

The two following results guarantee that for all $p \in M$, there exists a ball with large enough radius with center close to p that does not hit ∂M.

Lemma B.5. *Assume that $\partial M \neq \emptyset$. Let $q \in \partial M$ and $0 < t \leq \frac{\tau_M}{8} \wedge \frac{\tau_{\partial M}}{2}$. Then there exists $p_t \in \mathrm{Int}(M)$ such that*

- $\|p_t - q\| \in [t - 4t^2/\tau_M, t + 4t^2/\tau_M]$,
- $\mathrm{B}\big(p_t, t - 4t^2/\tau_M\big) \cap \partial M = \emptyset$.

Proof of Lemma B.5. Let η_q be the outward-pointing unit normal vector of M at q. Denote by $q_t := q - t\eta_q$, and $p_t := \pi_M(q_t)$. Note that $\mathrm{d}(q_t, M) \leq t < \tau_M$, so that p_t is well-defined.

Let us first prove that $p_t \notin \partial M$. For this, if we assume that $p_t \in \partial M$, then $p_t = \pi_{\partial M}(q_t)$ and, since $(q_t - q) \in N_q\partial M$ with $\|q_t - q\| < \tau_{\partial M}$, $p_t = \pi_{\partial M}(q_t) = q$. But as $p_t = q$, we get $\pi_M(q_t) = q$, with $\|q_t - q\| < \tau_M$. Thus, we conclude that $q_t - q = -t\eta_q \in Nor(q, M)$, which is a contradiction. Therefore, we do have $p_t \notin \partial M$ for $0 < t < \tau_M \wedge \tau_{\partial M}$.

Now, assume that $t \leq \frac{\tau_M}{8} \wedge \frac{\tau_{\partial M}}{2}$. For some unit vector $u_{p_t} \in (T_{p_t}M)^\perp$, it holds

$$\|p_t - q_t\| = \langle p_t - q_t, u_{p_t} \rangle .$$

Since $\|q_t - q\| = t \leq \tau_M/2$, [27, Theorem 4.8 (8)] entails that $\|p_t - q\| = \|\pi_M(q_t) - \pi_M(q)\| \leq \tau_M t/(\tau_M - t) \leq 2t$. From Proposition A.2, we deduce that

$$\angle(T_{p_t}M^\perp, T_qM^\perp) = \angle(T_{p_t}M, T_qM) \leq 4t/\tau_M.$$

Hence, there exists $u_q \in (T_qM)^\perp$ such that $\|u_q - u_{p_t}\| \leq 4t/\tau_M$. It follows that

$$\|p_t - q_t\| \leq \langle p_t - q_t, u_q \rangle + \frac{4t}{\tau_M} \|p_t - q_t\| ,$$

and thus, since $\eta_q \in T_qM$ and $u_q \in (T_qM)^\perp \subset Nor(q, M)$, we can write

$$\begin{aligned}
\frac{1}{2}\|p_t - q_t\| &\leq \left(1 - \frac{4t}{\tau_M}\right) \|p_t - q_t\| \\
&\leq \langle p_t - q_t, u_q \rangle \\
&= \langle p_t - q - t\eta_q, u_q \rangle \\
&= \langle p_t - q, u_q \rangle \\
&\leq \frac{\|p_t - q\|^2}{2\tau_M} \\
&\leq \frac{2t^2}{\tau_M},
\end{aligned}$$

where the last but one inequality follows from [27, Theorem 4.18]. As $\|q_t - q\| = t$, triangle inequality then yields $\|p_t - q\| \in [t - \frac{4t^2}{\tau_M}, t + \frac{4t^2}{\tau_M}]$. At last, since $\eta_q \in (T_q\partial M)^\perp$ and $t < \tau_{\partial M}$, $\mathring{\mathrm{B}}(q_t, t) \cap \partial M = \emptyset$. Noting that $\mathrm{B}(p_t, t - \frac{4t^2}{\tau_M}) \subset \mathring{\mathrm{B}}(q_t, t)$ concludes the proof. □

Corollary B.6. *For all $r \leq \frac{\tau_M}{32} \wedge \frac{\tau_{\partial M}}{3}$ and $x \in M$, there exists $x' \in \mathrm{B}(x, 3r/4) \cap M$ such that $\mathrm{B}(x', r/4) \cap \partial M = \emptyset$.*

Proof of Corollary B.6. Let us write $\Delta := \mathrm{d}(x, \partial M)$, with the convention $\mathrm{d}(x, \emptyset) = +\infty$. If $\Delta > r/2$, then taking $x' := x$ gives the result directly. We shall now assume that $\Delta \leq r/2$. Denote by $q := \pi_{\partial M}(x)$ and $q_t := q - t\eta_q$, where $t > 0$ and η_q is the unit outward-pointing vector of M at q.

Write $v := \pi_{Tan(q,M)}(x - q)$. Since $x - q \in (T_q\partial M)^\perp$ and that $\pi_{(T_q\partial M)^\perp}(Tan(q, M)) = \mathbb{R}_-\eta_q$ (see Proposition 2.6), we can write $v = -\ell\eta_q$ for some $\ell \geq 0$. Thus, we may decompose

$$x - q = -\ell\eta_q + u,$$

with $u \in Nor(q, M)$ and $\|u\| = \mathrm{d}(x - q, Tan(q, M)) \leq \Delta^2/(2\tau_M)$, from [27, Theorem 4.18]. From this decomposition, reverse triangle inequality yields

$$\begin{aligned} |\ell - \Delta| &= |\, \|-\ell\eta_q\| - \|x - q\| \,| \\ &\leq \|u\| \\ &\leq \Delta^2/(2\tau_M). \end{aligned}$$

We hence deduce that $\|x - q_\Delta\| \leq |\ell - \Delta| + \|u\| \leq \Delta^2/\tau_M$.

Now, pick $x' := \pi_M(q_{\Delta + r/2})$. It is immediate that $\|q_{\Delta+r/2} - q_\Delta\| = r/2$. Then, following the proof of Lemma B.5, since $\Delta + \frac{r}{2} \leq \frac{3r}{2} < \frac{\tau_M}{2}$, it holds

$$\|q_{\Delta+r/2} - x'\| \leq \frac{4(\Delta + r/2)^2}{\tau_M}.$$

These bounds altogether lead to

$$\begin{aligned} \|x' - x\| &\leq \left\|x' - q_{\Delta+r/2}\right\| + \left\|q_{\Delta+r/2} - q_\Delta\right\| + \|q_\Delta - x\| \\ &\leq \frac{4(\Delta + \frac{r}{2})^2}{\tau_M} + \frac{r}{2} + \frac{\Delta^2}{\tau_M} \\ &\leq r\left(\frac{1}{8} + \frac{1}{2} + \frac{1}{128}\right) \leq \frac{3r}{4}. \end{aligned}$$

At last, since $\Delta + \frac{r}{2} \leq \tau_{\partial M}/2$ and $(\Delta + \frac{r}{2}) - \frac{4(\Delta+\frac{r}{2})^2}{\tau_M} \geq \frac{r}{2}(1 - 1/6) > r/4$, we have

$$\begin{aligned} \mathrm{B}\left(x', \frac{r}{4}\right) \cap \partial M &\subset \mathring{\mathrm{B}}\left(x', \left(\Delta + \frac{r}{2}\right) - \frac{4(\Delta + \frac{r}{2})^2}{\tau_M}\right) \cap \partial M \\ &\subset \mathring{\mathrm{B}}\left(q_{\Delta+\frac{r}{2}}, \Delta + \frac{r}{2}\right) \cap \partial M \\ &= \emptyset, \end{aligned}$$

which concludes the proof. □

Proposition A.3 (Normal Vector Stability)**.** *Let $M \in \mathcal{M}^{d,D}_{\tau_{\min}, \tau_{\partial,\min}}$. Then for all $p, q \in \partial M$ such that $\|p - q\| \leq (\tau_M \wedge \tau_{\partial M})/32$, we have*

$$\|\eta_p - \eta_q\| \leq 9\|p - q\|/(\tau_M \wedge \tau_{\partial M}).$$

Proof of Proposition A.3. Let $p, q \in \partial M$, with $\|p-q\| = \kappa(\tau_M \wedge \tau_{\partial M})$, where $\kappa \leq 1/32$. According to Proposition A.2 and Proposition B.1 (applied with M), there exists $u \in T_qM$ such that $\|\eta_p - u\| \leq 2\|p-q\|/\tau_M \leq 2\kappa$. Decompose u as

$$u = \alpha\eta_q + v_q,$$

where $v_q \in T_q\partial M$. We may bound $\|v_q\|$ as follows. Let $w_q \in T_q\partial M$ with $\|w_q\| = 1$ be fixed. Using Proposition A.2 and Proposition B.1 again (but applied with ∂M), let $w_p \in T_p\partial M$ be such that $\|w_p - w_q\| \leq 2\|p-q\|/\tau_{\partial M} \leq 2\kappa$. We may write

$$\begin{aligned}\langle w_q, v_q\rangle &= \langle w_q, u\rangle \\ &= \langle w_p + (w_q - w_p), \eta_p + (u - \eta_p)\rangle \\ &\leq \frac{4(1+\kappa)\|p-q\|}{\tau_M \wedge \tau_{\partial M}},\end{aligned}$$

so that $\|v_q\| \leq 4(1+\kappa)\|p-q\|/(\tau_M \wedge \tau_{\partial M})$.

Next, let us prove that $\alpha \geq 0$ by contradiction. For this, assume that $\alpha < 0$, and let $\Delta_0 = (\tau_M \wedge \tau_{\partial M})/8$. Proceeding as in the proof of Lemma B.5 yields that

$$\mathrm{d}(q + \alpha\Delta_0\eta_q, M) \leq \frac{4\alpha^2\Delta_0^2}{\tau_M} \leq \frac{\Delta_0}{2}.$$

On the other hand, since $\eta_p \in Nor(p, M)$, [27, Theorem 4.8 (12)] asserts that $\mathring{\mathrm{B}}(p+\Delta_0\eta_p, \Delta_0) \cap M = \emptyset$. But triangle inequality allows to write

$$\begin{aligned}&\mathring{\mathrm{B}}\big(q + \alpha\Delta_0\eta_q, \Delta_0(1 - 10\kappa - 4\kappa(\kappa+1)\big) \cap M \\ &\quad\subset \mathring{\mathrm{B}}\big(q + (p-q) + \Delta_0(\eta_p - u) + \Delta_0\alpha\eta_q + \Delta_0 v_q, \Delta_0\big) \cap M \\ &\quad= \mathring{\mathrm{B}}(p + \Delta_0\eta_p, \Delta_0) \cap M \\ &\quad= \emptyset,\end{aligned}$$

so that we get to

$$\mathrm{d}(q + \alpha\Delta_0\eta_q, M) \geq (1 - 10\kappa - 4\kappa(\kappa+1))\Delta_0 > \Delta_0/2,$$

which is the desired contradiction. Thus, we have proven that $\alpha \geq 0$. Next, note that

$$\begin{aligned}1 = \|\eta_p\| &\leq \|\eta_p - u\| + \|u\| \\ &\leq \alpha + \|v_q\| + 2\kappa,\end{aligned}$$

so that $\alpha \geq 1 - 2\kappa - 4\kappa(1+\kappa) \geq 1/2$. Further, we may write

$$\begin{aligned}(1-\alpha)^2 + 2\alpha\,(1 - \langle\eta_p, \eta_q\rangle) &= \|\eta_p - \alpha\eta_q\|^2 \\ &\leq (\|\eta_p - u\| + \|v_q\|)^2 \\ &\leq \left(\frac{2 + 4(1+\kappa)}{\tau_M \wedge \tau_{\partial M}}\right)^2 \|p-q\|^2,\end{aligned}$$

that leads to

$$\begin{aligned}\|\eta_p - \eta_q\|^2 = 2\,(1 - \langle\eta_p, \eta_q\rangle) &\leq \left(\frac{2 + 4(1+\kappa)}{\tau_M \wedge \tau_{\partial M}}\right)^2 \frac{\|p-q\|^2}{\alpha} \\ &\leq 2\left(\frac{2 + 4(1+\kappa)}{\tau_M \wedge \tau_{\partial M}}\right)^2 \|p-q\|^2,\end{aligned}$$

hence the result. □

B.2 Projections and normals

Proposition A.4 (Tangent and Normal Components of Increments)**.** *Let $x, y \in M$, and $T \in \mathbb{G}^{D,d}$ be such that $\angle(T_xM, T) \leq \theta$. Write $(x-y)^T$ and $(x-y)^\perp$ for the orthogonal projection of $x-y$ onto T and $T^\perp$ respectively. Then,*

$$\|(y-x)^\perp\| \leq \|y-x\| \left(\theta + \|y-x\| / (2\tau_{\min})\right),$$
$$\|(y-x)^T\| \geq \|y-x\| \left(1 - \theta - \|y-x\| / (2\tau_{\min})\right).$$

Proof of Proposition A.4. Let $(y-x)^{T_x}$ and $(y-x)^{\perp_x}$ be the orthogonal projections of $y-x$ onto T_xM and $(T_xM)^\perp$ respectively. Since $\angle(T_xM, T) \leq \theta$, we have

$$\begin{aligned}
\left\|(y-x)^\perp\right\| &\leq \left\|((y-x)^{\perp_x})^\perp\right\| + \left\|((y-x)^{T_x})^\perp\right\| \\
&\leq \left\|(y-x)^{\perp_x}\right\| + \theta \left\|(y-x)^{T_x}\right\| \\
&\leq \frac{\|y-x\|^2}{2\tau_{\min}} + \theta \|y-x\|,
\end{aligned}$$

where the last line comes from [27, Theorem 4.18]. This proves the first inequality. The second one follows from the first one and triangle inequality. □

We now move to the proof of Proposition A.5, which we split into two intermediate results.

Proposition A.5 (Normals from Tangent Spaces)**.** *Let $x \in \partial M$, and $T \in \mathbb{G}^{D,d}$ such that $\angle(T_xM, T) < 1$. Then $T \cap Nor(x, M)$ contains a unique unit vector η, and it satisfies*

$$\|\eta - \eta_x\| \leq \sqrt{2}\angle(T_xM, T).$$

Proof of Proposition A.5. This is a straightforward consequence of Proposition B.7 and Proposition B.8. □

The following two results imply Proposition A.5. First, Proposition B.7 ensures that estimates of tangent spaces at boundary points contain a normal vector to ∂M. Second, Proposition B.8 ensures that this normal vector is close to the unit outward-pointing vector at the considered boundary point.

Proposition B.7. *Assume that $\partial M \neq \emptyset$. Let $x \in \partial M$ and $T \in \mathbb{G}^{D,d}$ be such that $\angle(T_xM, T) < 1$. Then $T \cap Nor(x, M)$ is a half-line: it contains a unique unit vector η.*

Furthermore, if $y \in \partial M$ and $(y-x)^\eta$ denotes the orthogonal projection of $(y-x)$ onto $\mathrm{span}(\eta)$*, we have*

$$\|(y-x)^\eta\| \leq \frac{\|y-x\|^2}{2\tau_{\partial M}}.$$

Proof of Proposition B.7. Since $\angle(T_xM, T) < 1$, for all $z \in \mathbb{R}^D \setminus \{0\}$,

$$\left\|(\pi_T + \pi_{T_xM^\perp})(z)\right\| = \|z - (\pi_T - \pi_{T_xM})(z)\| \geq (1 - \angle(T_xM, T)) \|z\| > 0.$$

Hence, $\pi_T + \pi_{T_xM^\perp}$ has full rank, which means that $\mathbb{R}^D = T + T_xM^\perp \subset T + N_xM$. Furthermore, $\dim(T) + \dim(N_xM) = D+1$ entails that $T \cap N_xM = \mathbb{R}u$ for some $u \neq 0$. We may thus decompose u as $u = u^{t_x} + u^{\eta_x} + u^{\perp_x}$, where $u^{t_x} = \pi_{N_xM^\perp}(u)$, $u^{\perp_x} = \pi_{T_xM^\perp}(u)$, and $u^{\eta_x} = \pi_{N_xM \cap T_xM}(u)$. Since $u \in N_xM$, we have $u^{t_x} = 0$, and the angle bound $\angle(T_xM, T) < 1$ yields that $\|u^{\eta_x}\| \geq$

$\|u\|(1-\angle(T_xM,T))>0$. As a result, $\eta := \mathrm{sign}(\langle u,\eta_x\rangle)u$ provides us with the announced unique unit $\eta \in T\cap Nor(x,M)$.

Now, the fact that $\eta \in Nor(x,M)\subset (T_x\partial M)^\perp$ allows to write

$$\begin{aligned}\|(y-x)^\eta\| &= |\,\langle y-x,\eta\rangle\,| \\ &= |\left\langle \pi_{(T_x\partial M)^\perp}(y-x),\eta\right\rangle| \\ &\le \left\|\pi_{(T_x\partial M)^\perp}(y-x)\right\| \\ &\le \frac{\|y-x\|^2}{2\tau_{\partial M}},\end{aligned}$$

where the last inequality follows from the reach condition on ∂M and [27, Theorem 4.18]. □

Proposition B.8. *Assume that $\partial M\neq\emptyset$. Let $x\in\partial M$ and $T\in\mathbb{G}^{D,d}$ be such that $\angle(T_xM,T)\le\theta<1$. Write η for* the *unit vector of $Nor(x,M)\cap T$ (Proposition B.7). Then,*

$$\|\eta-\eta_x\|\le\sqrt{2}\theta.$$

Proof of Proposition B.8. Since $\eta\in Nor(x,M)$, $\eta^{t_x}=0$. Furthermore, the angle condition yields that $\|\eta^{\perp_x}\|\le\theta\|\eta\|$. We may thus decompose $\eta=\langle\eta,\eta_x\rangle\,\eta_x+\beta u$ for some unit $u\in(\eta_x)^\perp$ and $|\beta|\le\theta$. In particular, $|\,\langle\eta,\eta_x\rangle\,|\ge\sqrt{1-\theta^2}$. But since $\eta\in Nor(x,M)$, $\langle\eta,\eta_x\rangle\ge 0$, so that in fact, $\langle\eta,\eta_x\rangle\ge\sqrt{1-\theta^2}$. Finally, as η and η_x are both unit vectors, we get

$$\|\eta-\eta_x\|=\sqrt{2}\sqrt{1-\langle\eta,\eta_x\rangle}\le\sqrt{2}\sqrt{1-\sqrt{1-\theta^2}}\le\sqrt{2}\theta.\square$$

Next, we state a simple lemma that will be useful for describing boundary balls.

Lemma B.9. *Assume that $\partial M\neq\emptyset$. Let $r<\tau_{\min}$, $x\in\partial M$ and $u\in N_x\partial M$ be such that $\langle\eta_x,u\rangle\ge 0$. Then* $\mathrm{B}(x+ru,r)\cap M=\{x\}$

Proof of Lemma B.9. As $u\in N_x\partial M$ and $\langle\eta_x,u\rangle\ge 0$, Proposition 2.6 yields that $u\in Nor(x,M)$, so that [27, Theorem 4.8 (12)] asserts that x is the unique projection of $x+ru$ onto M. □

The following result provides a quantitative bound on the metric distortion induced by projecting M locally onto (approximate) tangent spaces.

Proposition B.10. *Let $x\in M$ and $T\in\mathbb{G}^{D,d}$ be such that $\angle(T_xM,T)\le\theta$. Then, for all $y,z\in M\cap\mathrm{B}(x,\tau_{\min}/4)$, we have*

$$(6/10-\theta)\,\|y-z\|\le\|\pi_T(y)-\pi_T(z)\|\le\|y-z\|\,.$$

In particular, if $\theta\le 1/2$, then $\pi_T: M\cap\mathrm{B}(x,\tau_{\min}/4)\to\pi_T(M\cap\mathrm{B}(x,\tau_{\min}/4))$ is a homeomorphism.

Proof of Proposition B.10. The right hand side inequality is straightforward, since π_T is an orthogonal projection. For the other inequality, combine Proposition A.2 and Proposition B.1 to get

$$\begin{aligned}\angle(T,T_yM) &\le \angle(T,T_xM)+\angle(T_xM,T_yM)\\ &\le \theta+\frac{\mathrm{d}_M(x,y)}{\tau_{\min}}\\ &\le \theta+\left(1+\frac{\|y-x\|^2}{20\tau_{\min}^2}\right)\frac{\|y-x\|}{\tau_{\min}}\\ &\le \theta+(1+1/320)\frac{\|y-x\|}{\tau_{\min}}.\end{aligned}$$

Thus, Proposition A.4 applied at y and z entails

$$\begin{aligned}\|\pi_T(y) - \pi_T(z)\| &\geq \left(1 - \{\theta + (1+1/320)\,\|y-x\|/\tau_{\min}\} - \frac{\|y-z\|}{2\tau_{\min}}\right)\|y-z\| \\ &\geq (6/10 - \theta)\,\|y-z\|\,,\end{aligned}$$

which concludes the proof. □

For $q \in M$, the following result characterizes the boundary of $\pi_T(M \cap \mathrm{B}(q,r) - q)$, when seen as a subset of $T \cong \mathbb{R}^d$.

Lemma B.11. *Let $0 \leq r \leq \tau_{\min}/16$. Then for all $q \in M$ and $T \in \mathbb{G}^{D,d}$ such that $\angle(T_qM, T) \leq \theta \leq 1/8$,*

$$\partial\pi_{q+T}\big(M \cap \mathrm{B}(q,r)\big) = \pi_{q+T}\big(\partial M \cap \mathrm{B}(q,r)\big) \cup \pi_{q+T}\big(M \cap \partial\mathrm{B}(q,r)\big).$$

Proof of Lemma B.11. As preliminary remarks, first note that since $M \cap \mathrm{B}(q,r)$ is compact and π_{q+T} is continuous, we have

$$\overline{\pi_{q+T}(M \cap \mathrm{B}(q,r))} = \pi_{q+T}(M \cap \mathrm{B}(q,r)).$$

Furthermore, for all $p \in \mathrm{B}(q,r)$, Proposition A.2 and Lemma A.1 yield that $\angle(T_pM, T) \leq 1/4$. We recall that $\mathrm{Int}(M) = M \setminus \partial M$.

Step 1: First, we prove that $\pi_{q+T}\big(\mathrm{Int}(M) \cap \mathring{\mathrm{B}}(q,r)\big) \subset \big(\pi_{q+T}(\mathrm{B}(q,r) \cap M)\big)^{\mathrm{o}}$.
For this, let $p \in \mathrm{Int}(M) \cap \mathring{\mathrm{B}}(q,r)$ be fixed. Let $\rho_M \in (0, \min\{r - \|p-q\|, \mathrm{d}(p,\partial M)\})$ (with the convention $\mathrm{d}(p,\emptyset) = +\infty$), so that in particular, $M \cap \mathring{\mathrm{B}}(p,\rho_M) \subset \mathrm{Int}(M) \cap \mathring{\mathrm{B}}(q,r)$. According to [3, Lemma 1], there exists $0 < r_2 \leq \tau_M/8$ such that

$$\exp_p : \mathring{\mathrm{B}}_{T_pM}(0,r_2) \longrightarrow \mathring{\mathrm{B}}(p,\rho_M) \cap \mathrm{Int}(M)$$

is a diffeomorphism onto its image, and can be decomposed as $\exp_p(v) = p + v + N_p(v)$, with $N_p(0) = 0$, $d_0N_p = 0$, $\|d_vN_p\|_{op} \leq 5/(4\tau_M)$. We now consider the map g defined as

$$\begin{aligned} g\colon \mathring{\mathrm{B}}_T(0,r_2) &\to \mathring{\mathrm{B}}(p,\rho_M) \cap \mathrm{Int}(M) \\ u &\mapsto \exp_p(\pi_{T_pM}(u))\end{aligned}$$

Note that, since $\angle(T_pM, T) \leq 1/4$, $\pi_{T_pM} : \mathring{\mathrm{B}}_T(0,r_2) \to \mathring{\mathrm{B}}_{T_pM}(0,r_2)$ is a diffeomorphism onto its image that satisfies $\|u - \pi_{T_pM}(u)\| \leq \|u\|/4$ for all $u \in \mathring{\mathrm{B}}_T(0,r_2)$. In particular, π_{T_pM} is injective on T, and hence so is g on its domain. As a result, for all $u_1, u_2 \in \mathring{\mathrm{B}}_T(0,r_2)$,

$$\begin{aligned} g(u_1) - g(u_2) = (u_1 - u_2) &+ (\pi_{T_pM}(u_1 - u_2) - (u_1 - u_2)) \\ &+ N_p(\pi_{T_pM}(u_1)) - N_p(\pi_{T_pM}(u_2)).\end{aligned}$$

We may thus bound

$$\begin{aligned}\|g(u_1) - g(u_2) - (u_1 - u_2)\| &\leq \frac{1}{4}\|u_1 - u_2\| + 5r_2/(4\tau_{\min})\|u_1 - u_2\| \\ &\leq \frac{1}{2}\|u_1 - u_2\|.\end{aligned}$$

Let now $f : \mathring{\mathrm{B}}_T(0, r_2) \to \mathring{\mathrm{B}}_T(0, \rho_M)$ be defined as $f(\cdot) := \pi_{q+T} \circ (g(\cdot) - p)$. By composition and Proposition B.10, f is clearly injective. Moreover, for all $u_1, u_2 \in \mathring{\mathrm{B}}_T(0, r_2)$,

$$\frac{1}{2}\|u_1 - u_2\| \leq \|f(u_1) - f(u_2)\| \leq \frac{3}{2}\|u_1 - u_2\|,$$

since $\pi_T(u_1 - u_2) = u_1 - u_2$ and $\|\pi_T\left(g(u_1) - g(u_2) - (u_1 - u_2)\right)\| \leq \|u_1 - u_2\|/2$. Thus, $f : \mathring{\mathrm{B}}_T(0, r_2) \to f(\mathring{\mathrm{B}}_T(0, r_2))$ is a homeomorphism, which ensures that $f(\mathring{\mathrm{B}}_T(0, r_2))$ is an open subset of T that contains $0 = f(0)$. But by construction,

$$\pi_{q+T}(p) + f(\mathring{\mathrm{B}}_T(0, r_2)) \subset \pi_{q+T}(\mathring{\mathrm{B}}(p, \rho_M) \cap \mathrm{Int}(M)),$$

which shows that $\pi_{q+T}(p) \in (\pi_{q+T}(\mathrm{B}(q, r) \cap M))^{\mathrm{o}}$, and concludes the first step.

Step 2: Next, we show that no element of $\pi_{q+T}((\partial M \cap \mathrm{B}(q, r)) \cup (M \cap \mathcal{S}(q, r)))$ can belong to the interior set $(\pi_{q+T}(\mathrm{B}(q, r) \cap M))^{\mathrm{o}}$.

– If $\partial M \neq \emptyset$, let $p \in \partial M \cap \mathrm{B}(q, r)$ be fixed. Striving for a contradiction, assume that $\pi_{q+T}(p) \in \pi_{q+T}(M \cap \mathrm{B}(q, r))^{\mathrm{o}}$. In particular, for $\delta > 0$ small enough, $\pi_{q+T}(p + \delta\eta_p) \in \pi_{q+T}(\mathrm{B}(q, r) \cap M)$. Without loss of generality, we shall pick $\delta \in (0, \tau_{\min}/16)$ small enough so that $p + \delta\eta_p \in \mathrm{B}(q, r)$. Then there exists $p' \in \mathrm{B}(q, r) \cap M$ such that $\pi_{q+T}(p') = \pi_{q+T}(p + \delta\eta_p)$, or equivalently, $\pi_T(p' - p) = \delta\pi_T(\eta_p)$. Consider $v := p' - p - \delta\eta_p$. By construction, $\pi_T(v) = 0$, so that $v \in T^{\perp}$, and its norm is at most

$$\|v\| \leq \left\|p' - p\right\| + \|\delta\eta_p\| \leq 2r + \delta \leq 3\tau_{\min}/8.$$

Furthermore, $v \neq 0$, as otherwise this would mean that $p + \delta\eta_p = p' \in \mathrm{B}(q, r) \cap M \subset M$, which is impossible since $\mathrm{d}(p + \delta\eta_p, M) = \delta$ from [27, Theorem 4.8 (12)]. We may now decompose v as $v = v_1 + v_2$, with $v_1 \in T_pM$ and $v_2 \in T_pM^{\perp}$.

∗ On one hand, the angle bound $\angle(T, T_pM) \leq 1/4$ and $v \in T_pM^{\perp}$ yield $\|v_1\| \leq \|v\|/4$.

∗ Furthermore, $\delta \leq \tau_{\min}/16$ ensures that $\|v_2\| \leq \|v\| \leq 3\tau_{\min}/8 < \tau_M - \delta$. Let us now consider $s := p + \delta\eta_p + v_2$. As $\delta\eta_p + v_2 \in Nor(p, M)$ and $\|\delta\eta_p + v_2\| < \tau_M$, [27, Theorem 4.8 (12)] asserts that $\pi_M(s) = p$ and $\mathrm{d}(s, M) = \|\delta\eta_p + v_2\|$. But on the other hand, $s + v_1 = p' \in M$, so clearly $\|v_1\| \geq \mathrm{d}(s, M)$. Therefore,

$$\begin{aligned} \|v_1\|^2 &\geq \|\delta\eta_p + v_2\|^2 \\ &= \delta^2 + \|v_2\|^2 \\ &= \delta^2 + \|v\|^2 - \|v_1\|^2 \\ &\geq \|v\|^2 - \|v_1\|^2, \end{aligned}$$

and thus $\|v_1\| \geq \|v\|/\sqrt{2}$.

The last two items contradicting each other, we finally obtain that $p \notin \pi_{q+T}(M \cap \mathrm{B}(q, r))^{\mathrm{o}}$.

– Let now $p \in \partial\mathrm{B}(q, r) \cap M$ be fixed. Striving for a contradiction, let us assume that $\pi_{q+T}(p) \in \pi_{q+T}(M \cap \mathrm{B}(q, r))^{\mathrm{o}}$. This implies in particular that for all $\delta < 1$ small enough, $\pi_{q+T}(p + \delta(p - q)) \in \pi_{q+T}(\mathrm{B}(q, r) \cap M)$. Then there exists $v \in T^{\perp}$ such that $p + \delta(p - q) + v \in M \cap \mathrm{B}(q, r)$. Denote by $v_2 = \pi_{T_pM^{\perp}}(v)$. Since $\angle(T_pM, T) \leq 1/4$, we have $\|v\| \geq 3\|v_2\|/4$. On the other

hand, since $p+\delta(q-p)+v \in M$, we have

$$\begin{aligned}\|\pi_{T_pM^\perp}(\delta(p-q)+v)\| &= \mathrm{d}\big((p+\delta(p-q)+v)-p, T_pM\big)\\ &\leq \frac{\|\delta(p-q)+v\|^2}{2\tau_M}\\ &\leq \frac{\delta^2 r^2+\|v\|^2}{\tau_M},\end{aligned}$$

from [27, Theorem 4.18]. And noting that

$$\begin{aligned}\Big\|\pi_{T_pM^\perp}(\delta(p-q)+v)\Big\| &= \Big\|\delta\pi_{T_pM^\perp}(p-q)+v_2\Big\|\\ &\geq \|v_2\|-\delta\mathrm{d}(q-p,T_pM)\\ &\geq \frac{3\|v\|}{4}-\frac{\delta r^2}{2\tau_M},\end{aligned}$$

we obtain

$$\|v\| \leq \frac{4}{3}\left(\frac{\delta r^2}{2\tau_M}+\frac{\delta^2 r^2+\|v\|^2}{\tau_M}\right) \leq 2\left(\frac{\delta r^2}{2\tau_M}+\frac{\delta^2 r^2+\|v\|^2}{\tau_M}\right). \tag{13}$$

On the other hand, since $p+\delta(p-q)+v \in \mathrm{B}(q,r)$, we have $\|(1+\delta)(p-q)+v\|^2 \leq r^2$, and therefore

$$(2\delta+\delta^2)r^2+\|v\|^2-2(1+\delta)r\|v\| \leq 0,$$

But according to (13), this last inequality yields

$$\begin{aligned}&(2\delta+\delta^2)r^2+\|v\|^2-2(1+\delta)r\|v\|\\ &\qquad\geq (2\delta+\delta^2)r^2+\|v\|^2-4(1+\delta)r\left(\frac{\delta r^2}{2\tau_M}+\frac{\delta^2 r^2+\|v\|^2}{\tau_M}\right)\\ &\qquad= \|v\|^2\left(1-4(1+\delta)\frac{r}{\tau_M}\right)+r^2\left((2\delta+\delta^2)-4(1+\delta)\left\{\frac{\delta r}{2\tau_M}+\frac{r\delta^2}{\tau_M}\right\}\right),\end{aligned}$$

and since $r \leq \tau_M/16$ and $\delta \in (0,1]$, we finally get

$$\begin{aligned}(2\delta+\delta^2)r^2+\|v\|^2-2(1+\delta)r\|v\| &\geq \frac{\|v\|^2}{2}+r^2\left((2\delta+\delta^2)-\delta(1+\delta)\left\{\frac{1}{8}+\frac{1}{4}\right\}\right)\\ &\geq \frac{\|v\|^2}{2}+r^2\delta\\ &> 0\end{aligned}$$

which is the desired contradiction. That is, we have $\pi_{q+T}(p) \notin \pi_{q+T}(M \cap \mathrm{B}(q,r))^{\mathrm{o}}$, as announced.

Conclusion: Putting everything together, we deduce that

$$\begin{aligned}\pi_{q+T}((\partial M \cap \mathrm{B}(q,r)) \cup (M \cap \partial\mathrm{B}(q,r))) &= \overline{\pi_{q+T}(M \cap \mathrm{B}(q,r)} \setminus \pi_{q+T}(M \cap \mathrm{B}(q,r))^{\mathrm{o}}\\ &= \partial\pi_{q+T}(M \cap \mathrm{B}(q,r),\end{aligned}$$

which is the announced result.

□

B.3 Structure of balls on manifolds with boundary

Using Lemma B.11, we are now able to derive the two key results on the structure of $\pi_T(\mathrm{B}(x, R_0) - x)$. This structure depends on whether x is either near or far from ∂M. We start with the case where x is an interior point.

Lemma A.6 (Far-Boundary Balls)**.** *Let $x \in M$ and $T \in \mathbb{G}^{D,d}$ be such that $\angle(T_x M, T) \leq \theta \leq 1/8$. If $\mathrm{d}(x, \partial M) > 0$ (with the convention $\mathrm{d}(x, \emptyset) = +\infty$), and $R \leq \tau_{\min}/16$, then*

$$\mathrm{B}_T\left(0, \frac{4}{5}\min\left\{R, \mathrm{d}(x, \partial M)\right\}\right) \subset \pi_T(\mathrm{B}(x, R) \cap M - x).$$

Proof of Lemma A.6. Let z' be in $\mathring{\mathrm{B}}\left(x, 4\min\left\{R, \mathrm{d}(x, \partial M)\right\}/5\right) \cap (x + T)$, and assume for contradiction that $z' \notin \pi_{x+T}(\mathrm{B}(x, R) \cap M)$. Then by connectedness, there exists $z \in [x, z']$ such that $z \in \partial \pi_{x+T}(\mathrm{B}(x, R) \cap M)$.

- Note that, since $\mathring{\mathrm{B}}\left(x, 4\min\left\{R, \mathrm{d}(x, \partial M)\right\}/5\right) \cap (x + T)$ is convex and contains $\{x, z'\}$, we have $z \in \mathring{\mathrm{B}}\left(x, 4\min\left\{R, \mathrm{d}(x, \partial M)\right\}/5\right) \cap x + T$.
- According to Lemma B.11, we can write $z = \pi_{x+T}(y)$ with $y \in \partial\mathrm{B}(x, R) \cap M$ or $y \in \mathrm{B}(x, R) \cap \partial M$. Therefore, we have either $\|y - x\| = R$ or $\|y - x\| \geq \mathrm{d}(x, \partial M)$, which entails $\|y - x\| \geq \min\left\{R, \mathrm{d}(x, \partial M)\right\}$. Applying Proposition A.4 gives that

$$\begin{aligned}
\|x - z\| &= \|\pi_T(x) - \pi_T(z)\| \\
&\geq \min\left\{R, \mathrm{d}(x, \partial M)\right\}\left(1 - \theta - \frac{\|x - y\|}{2\tau_{\min}}\right) \\
&\geq \frac{27}{32}\min\left\{R, \mathrm{d}(x, \partial M)\right\} \\
&\geq \frac{4}{5}\min\left\{R, \mathrm{d}(x, \partial M)\right\},
\end{aligned}$$

  leading to $z \notin \mathring{\mathrm{B}}\left(x, 4\min\left\{R, \mathrm{d}(x, \partial M)\right\}/5\right)$, and hence a contradiction.

It follows that $\mathring{\mathrm{B}}\left(x, 4\left\{R, \mathrm{d}(x, \partial M)\right\}/5\right) \cap (x + T) \subset \pi_{x+T}(\mathrm{B}(x, R) \cap M)$. Finally, the closedness of $\pi_{x+T}(\mathrm{B}(x, R) \cap M)$ concludes the proof. □

Next we turn to the case where x is a boundary point.

Lemma A.7 (Near-Boundary Balls)**.** *Assume that $\partial M \neq \emptyset$. Let $x \in \partial M$ and $T \in \mathbb{G}^{D,d}$ be such that $\angle(T_x M, T) \leq \theta \leq 1/8$. Denote by $\hat{\eta}$ the unit vector of $T \cap Nor(x, M)$, choose $R \leq \tau_{\min}/16$ and $r \leq \min\left\{2R/5, 7\tau_{\partial,\min}/5\right\}$.*

Then, writing $O^{in} := -r\hat{\eta}$ and $O^{out} := r\hat{\eta}$, we have

$$\mathrm{B}(O^{in}, r) \cap T \subset \pi_T(\mathrm{B}(x, R) \cap M - x) \subset \mathring{\mathrm{B}}(O^{out}, r)^c \cap T.$$

Proof of Lemma A.7. Take $O = x + \alpha\hat{\eta}$ with $|\alpha| = r$.

We first prove that $(\mathrm{B}(O, r) \cap (x + T)) \cap \partial\pi_{x+T}(\mathrm{B}(x, R) \cap M) = \{x\}$. For this, consider $z \in \pi_{x+T}(M \cap \mathrm{B}(x, R)) \setminus \{x\}$ and $y \in M \cap \mathrm{B}(x, R)$ such that $z = x + (y-x)^T = x + (y-x)^{\hat{t}} + (y-x)^{\hat{\eta}}$. Recall that $(y - x)^{\hat{t}}$ denotes the orthogonal projection of $y - x$ onto $\hat{\eta}^\perp \cap T$. We have that

$$\begin{aligned}
\|O - z\|^2 &= \left(\left\|(y-x)^{\hat{\eta}}\right\| \pm |\alpha|\right)^2 + \left\|(y-x)^{\hat{t}}\right\|^2 \\
&\geq \left(\left\|(y-x)^{\hat{\eta}}\right\| - |\alpha|\right)^2 + \left\|(y-x)^{\hat{t}}\right\|^2 \\
&= r^2 + \left\|(y-x)^T\right\|^2 - 2r\left\|(y-x)^{\hat{\eta}}\right\|.
\end{aligned}$$

According to Lemma B.11, if $z \in \partial\pi_{x+T}(M \cap \mathrm{B}(x,R))$, we have either $z \in \pi_{x+T}(M \cap \partial\mathrm{B}(x,R))$, or $z \in \pi_{x+T}(\partial M \cap \mathrm{B}(x,R))$. In the first case, Proposition A.4 gives

$$\begin{aligned}\|O-z\|^2 &\geq r^2 + \left\|(y-x)^T\right\|^2 - 2r\left\|(y-x)^T\right\| \\ &\geq r^2 + \left\|(y-x)^T\right\|\left(\left\|(y-x)^T\right\| - 2r\right) \\ &\geq r^2 + \left\|(y-x)^T\right\|\left(\frac{27}{32}R - 2r\right).\end{aligned}$$

In the second case, using Proposition A.4 and Proposition B.7 leads to

$$\|O-z\|^2 \geq r^2 + \|y-x\|^2\left(\left(\frac{27}{32}\right)^2 - \frac{r}{2\tau_{\partial,\min}}\right).$$

In both cases, since $z \neq x$ by assumption, we have $(y-x)^T \neq 0$ and hence $y-x \neq 0$, so that if $r \leq \min\{2R/5, 7\tau_{\partial,\min}/5\}$, we have $\|O-z\| > r$, which entails $z \notin \mathrm{B}(O,r)$. In other words, we have proved that $\mathrm{B}(O,r) \cap \partial\pi_{x+T}(\mathrm{B}(x,R_0)\cap M) = \{x\}$.

By connectedness, it follows that if $O \in \{x+O^{\mathrm{in}}, x+O^{\mathrm{out}}\}$, we have either

$$\mathrm{B}(O,r) \cap (x+T) \subset \pi_{x+T}(\mathrm{B}(x,R_0)\cap M),$$

or

$$\mathrm{B}(O,r) \cap (x+T) \subset (\pi_{x+T}(\mathrm{B}(x,R_0)\cap M)^c \cup \{x\}.$$

Let us now focus on $\mathrm{B}(x+O^{\mathrm{out}},r)\cap(x+T)$. Consider a sequence $x_n^* = x+\varepsilon_n\hat{\eta}$ with $\varepsilon_n > 0$ converging to 0. Suppose that $x_n^* \in \pi_{x+T}(\mathrm{B}(x,R)\cap M)$ i.e. there exits $x_n \in M$ such that $x_n^* - x = (x_n-x)^T$. By Proposition A.4, we have $\left\|(x_n-x)^\perp\right\| \leq \varepsilon_n(\theta+1/4)$. Let $\Omega = x+r'\hat{\eta}$ with $r' < \min(\tau_{\min}, \tau_{\partial,\min})$. On one hand Lemma B.9 ensures that $\|\Omega - x_n\| \geq r'$ and, on the other hand

$$\|\Omega-x_n\|^2 = (r'-\varepsilon_n)^2 + \left\|(x_n-x)^\perp\right\|^2 \leq r'^2 - 2\varepsilon_n r' + \varepsilon_n^2\left(1+(\theta+1/4)^2\right).$$

Thus, for n large enough $\|\Omega-x_n\|^2 < r'^2$, which is impossible. Hence, for n large enough $x_n^* \notin \pi_{x+T}(\mathrm{B}(x,R)\cap M)$, which proves the right hand side inclusion

$$\pi_{x+T}(\mathrm{B}(x,R)\cap M) \subset \left(\mathrm{B}(x+O^{\mathrm{out}},r)^c \cap (x+T)\right) \cup \{x\}.$$

Next, we prove that if $\theta \leq 1/8$, then there exists $x^* \in x+T\cap\mathrm{B}(x+O^{\mathrm{in}},r)$ such that $x^* \in \pi_{x+T}(\mathrm{B}(x,R)\cap M)$, and thus $\mathrm{B}(x+O^{\mathrm{in}},r)\cap x+T \subset \pi_{x+T}(\mathrm{B}(x,R_0)\cap M)$. For this, introduce $\eta = \pi_{T_xM}(\hat{\eta})$ and $\eta' = \pi_T(\eta)$. We clearly have $\|\eta\| \leq 1$, $\|\eta'\| \leq 1$, $\|\hat{\eta}-\eta\| \leq \theta$ and $\|\eta-\eta'\| \leq \theta$. In particular, this implies that $\|\eta'-\hat{\eta}\| \leq 2\theta < 1$ and $\|\eta'\| \geq 1-2\theta$. Hence, decomposing $\eta' = \lambda\hat{\eta}+\mu v$, with $v \in T\cap(\hat{\eta})^\perp$ and $\|v\|=1$, we have $\lambda > 0$, with

$$(1-2\theta)^2 \leq \lambda^2+\mu^2 \leq 1 \text{ and } \lambda \geq 1-2\theta.$$

Furthermore, since $\eta \in T_xM$ and that

$$\langle \eta, \eta_x\rangle \geq 1-\|\eta-\eta_x\| \geq 1-\|\eta_x-\hat{\eta}\|-\|\hat{\eta}-\eta\| \geq 1-\sqrt{2}\theta-\theta > 0$$

from Proposition B.8, we get that $\eta \in Nor(x,M)$ from Proposition 2.6, or equivalently that $-\eta \in Tan(x,M)$. Hence, [27, Definition 4.3] asserts that there exists a sequence $(x_n)_n \in M\setminus\{x\}$ converging to x such that $\left\|\frac{x_n-x}{\|x_n-x\|} - \frac{-\eta}{\|\eta\|}\right\| \leq \frac{1}{n}$, that is

$$x_n = x - \|x-x_n\|\left(\frac{\eta}{\|\eta\|} + \frac{1}{n}w_n\right) \text{ with } \|w_n\| \leq 1.$$

Considering $x_n^* = \pi_{x+T}(x_n)$, $w_n^* = \pi_T(w_n)$, and $\varepsilon_n = \frac{\|x-x_n\|}{\|\eta\|}$, we may hence write

$$x_n^* = x - \varepsilon_n \left(\lambda\hat{\eta} + \mu v + \frac{\|\eta\|}{n} w_n^* \right),$$

so that

$$\begin{aligned} \left\| x + O^{\text{in}} - x_n^* \right\| &\leq \|(r - \lambda\varepsilon_n)\hat{\eta} + \varepsilon_n \mu v\| + \frac{\varepsilon_n}{n} \\ &\leq \sqrt{r^2 - 2r\lambda\varepsilon_n + \varepsilon_n^2} + \frac{\varepsilon_n}{n} \\ &\leq \sqrt{(r - \lambda\varepsilon_n)^2 + \varepsilon_n^2(1-\lambda^2)} + \frac{\varepsilon_n}{n} \\ &\leq (r - \lambda\varepsilon_n) + \varepsilon_n\sqrt{1-\lambda^2} + \frac{\varepsilon_n}{n}. \end{aligned}$$

Since $\lambda \geq 1 - 2\theta \geq 3/4$, this yields

$$\left\| x + O^{\text{in}} - x_n^* \right\| \leq r - \varepsilon_n \left(\frac{3}{4} - \frac{\sqrt{7}}{4} + \frac{1}{n} \right).$$

On the other hand, we have

$$\|x_n^* - x\| \geq \frac{\|x_n - x\|}{\|\eta\|} \left(\sqrt{\lambda^2 + \mu^2} - \frac{1}{n} \right) \geq \frac{\|x_n - x\|}{\|\eta\|} \left(\frac{3}{4} - \frac{1}{n} \right) > 0,$$

for n large enough. Thus, for n large enough, $x_n^* \in (x+T) \cap \mathrm{B}(x + O^{\text{in}}, r)$ with $x_n^* \in \pi_{x+T}(\mathrm{B}(x, R) \cap M)$ and $x_n^* \neq x$, ensuring that

$$\mathrm{B}(x + O^{\text{in}}, r) \cap (x + T) \subset \pi_{x+T}(\mathrm{B}(x, R_0) \cap M),$$

which is the left hand side inclusion. □

At last, the following consequence of Lemma A.7 will be of particular interest in the proof of Theorem 5.1.

Corollary A.8 (Parallelism of Projected Normals)**.** *Assume that $\partial M \neq \emptyset$. Let $x \in M$ be such that $\mathrm{d}(x, \partial M) < \tau_{\min}/16$, and $y \in \mathbb{R}^D$. For $T \in \mathbb{G}^{D,d}$, let $x^* \in \pi_{y+T}(\partial M \cap \mathrm{B}(x, \tau_{\min}/16))$ be any point such that*

$$\|x^* - \pi_{y+T}(x)\| = \mathrm{d}(\pi_{y+T}(x), \pi_{y+T}(\partial M \cap \mathrm{B}(x, \tau_{\min}/16)),$$

and

$$x' \in \partial M \cap \mathrm{B}(x, \tau_{\min}/16) \text{ such that } \pi_{y+T}(x') = x^*.$$

If $\angle(T_{x'}M, T) \leq 1/8$, then $Nor(x', M) \cap T$ contains a unique unit vector $\eta^*(x')$, and

$$x^* - \pi_{y+T}(x) = \|x^* - \pi_{y+T}(x)\|\, \eta^*(x').$$

Proof of Corollary A.8. According to Proposition A.5, $Nor(x', M) \cap T$ contains a unique unit vector $\eta^*(x')$. By definition of x^* we have

$$\mathring{\mathrm{B}}_{y+T}(\pi_{y+T}(x), \|x^* - \pi_{y+T}(x)\|) \cap \pi_{y+T}(\partial M \cap \mathrm{B}(x, \tau_{\min}/16)) = \emptyset. \tag{14}$$

Since $\pi_{y+T} = \pi_{x'+T} + \pi_{T^\perp}(y - x')$, Lemma A.7 applied at x' with $R_0 = \tau_{\min}/16$ yields

$$\mathring{\mathrm{B}}_{x'+T}(x' + r_0\eta^*(x'), r_0) \cap \pi_{x'+T}(M \cap \mathrm{B}(x', R_0)) = \emptyset.$$

Since $\pi_{x'+T} = \pi_{y+T} + \pi_{T^\perp}(x' - y)$, and that for all $p \in x' + T$ and $r > 0$,

$$\mathring{\mathrm{B}}_{x'+T}(p, r) = \pi_{T^\perp}(x' - y) + \mathring{\mathrm{B}}_{y+T}(\pi_{y+T}(p), r),$$

we deduce that

$$\mathring{\mathrm{B}}_{y+T}(x^* + r_0\eta^*(x'), r_0) \cap \pi_{y+T}(M \cap \mathrm{B}(x', R_0)) = \emptyset. \tag{15}$$

Now, decompose

$$x^* - \pi_{y+T}(x) = \cos\varphi \, \|x^* - \pi_{y+T}(x)\| \, \eta^*(x') + \sin\varphi \, \|x^* - \pi_{y+T}(x)\| \, v$$

with $v \in \eta^*(x')^\perp$ and $\varphi \in [0, 2\pi)$, and consider

$$x_t := x^* + t\sin(\pi - \varphi/2)\eta^*(x') + t\cos(\pi - \varphi/2)v,$$

for $t \geq 0$. Straightforward calculus yields

$$\begin{cases} \|x^* + r_0\eta^*(x') - x_t\|^2 = r_0^2 + t^2 - 2r_0 t \sin(\pi - \varphi/2), \\ \|\pi_{y+T}(x) - x_t\|^2 = \|x^* - \pi_{y+T}(x)\|^2 + t^2 + 2t\,\|x^* - \pi_{y+T}(x)\| \sin(\pi + \varphi/2), \\ \|x - x_t\| \leq \|x - x^*\| + t \text{ with } \|x - x^*\| \leq \mathrm{d}(x, \partial M) < \tau_{\min}/16. \end{cases}$$

Suppose, to derive a contradiction, that $\varphi \neq 0$. Then for small enough t, we have

$$x_t \in \mathring{\mathrm{B}}(x, \tau_{\min}/16) \cap \mathring{\mathrm{B}}_{y+T}(x^* + r_0\eta^*(x'), r_0) \cap \mathring{\mathrm{B}}_{y+T}(\pi_{y+T}(x), \|x^* - \pi_{y+T}(x)\|).$$

Then, Equation (15) provides $z \in (x_t, \pi_{y+T}(x))$ such that $z \in \pi_{y+T}(\partial M \cap \mathrm{B}(x, \tau_{\min}/16))$. But since $\|z - \pi_{y+T}(x)\| < \|x^* - \pi_{y+T}(x)\|$ by construction, Equation (14) leads to the desired contradiction. Hence, $\varphi = 0$, which yields the announced result. □

B.4 Volume bounds and covering numbers

Lemma A.9 (Sampling Density Bound)**.** *Let $\varepsilon_1 = \left(C_d \frac{\log n}{f_{\min} n}\right)^{\frac{1}{d}}$, for C_d large enough. Then, for n large enough so that $\varepsilon_1 \leq \frac{\tau_{\min}}{16} \wedge \frac{\tau_{\partial,\min}}{2}$, we have, with probability larger than $1 - n^{-3}$,*

$$\mathrm{d_H}(M, \mathbb{X}_n) \leq \varepsilon_1.$$

Proof of Lemma A.9. Let $\varepsilon_1 \leq \frac{\tau_{\min}}{16} \wedge \frac{\tau_{\partial,\min}}{2}$, and $x \in M$. As $\mathbb{X}_n \subset M$, the Hausdorff distance between M and $\mathbb{X}_n$ writes as $\mathrm{d_H}(M, \mathbb{X}_n) = \max_{x \in M} \mathrm{d}(x, \mathbb{X}_n)$. Furthermore, according to Corollary B.6,

$$\begin{aligned} \mathbb{P}\left(\max_{x \in M} \mathrm{d}(x, \mathbb{X}_n) \geq \varepsilon_1\right) &\leq \mathbb{P}\left(\max_{\substack{x' \in M \\ \mathrm{d}(x', \partial M) \geq \varepsilon_1/4}} \mathrm{d}(x', \mathbb{X}_n) \geq \varepsilon_1/4\right) \\ &\leq \frac{16^d}{c_d f_{\min} \varepsilon_1^d} \exp\left(-n \frac{c_d f_{\min}}{8^d} \varepsilon_1^d\right), \end{aligned}$$

where the second inequality follows as [2, Lemma 9.1]. Thus, choosing $\varepsilon_1 = \left(C_d \frac{\log n}{f_{\min} n}\right)^{\frac{1}{d}}$, for C_d large enough, yields that $\mathrm{d_H}(M, \mathbb{X}_n) \leq \varepsilon_1$, with probability larger than $1 - n^{-3}$. □

Lemma B.12 (Volume of Intersection of Balls)**.** *Let $0 \le r' \le r$, and $O, O' \in \mathbb{R}^d$ that satisfy*

$$\|O - O'\| = r + r' - h,$$

for some $0 \le h \le r'$. Then

$$\mathcal{H}^d\big(\mathrm{B}(O,r) \cap \mathrm{B}(O',r')\big) \ge \frac{\omega_{d-1}}{d 2^{\frac{d-1}{2}}} h^{\frac{d+1}{2}} (r')^{\frac{d-1}{2}}.$$

Proof of Lemma B.12. Let $A := \partial\mathrm{B}(O,r) \cap [O,O']$, $B := \partial\mathrm{B}(O',r') \cap [O,O']$, and Ω be the orthogonal projection of any point of $\partial\mathrm{B}(O,r) \cap \partial\mathrm{B}(O',r')$ onto $[O,O']$. Also define $a := \|A - \Omega\|$, $b := \|B - \Omega\|$ and $\ell := \mathrm{d}(\Omega, \partial\mathrm{B}(O,r) \cap \partial\mathrm{B}(O',r'))$ (see Figure 9). Let $\mathcal{C}$ (resp. $\mathcal{C}'$) denote the section of cone of apex B (resp. A), direction $O - O'$ (resp. $O' - O$), and basis $\mathrm{B}(\Omega,\ell) \cap \big(\Omega + \mathrm{span}(O' - O)^\perp\big)$.

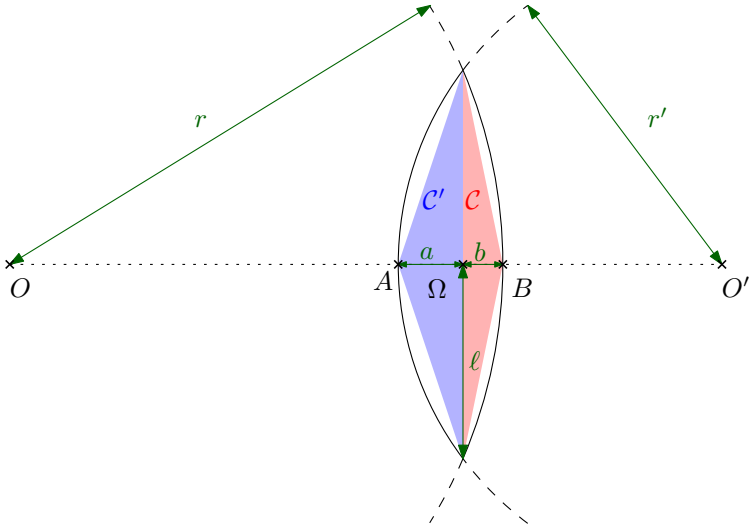


Figure 9: Layout for Lemma B.12.

By convexity, we have $\mathcal{C}, \mathcal{C}' \subset \mathrm{B}(O,r) \cap \mathrm{B}(O',r')$, and since $\mathcal{C} \cap \mathcal{C}'$ is included in a hyperplane, we get

$$\begin{aligned}
\mathcal{H}^d\big(\mathrm{B}(O,r) \cap \mathrm{B}(O',r')\big) &\ge \mathcal{H}^d(\mathcal{C} \cup \mathcal{C}') \\
&= \mathcal{H}^d(\mathcal{C}) + \mathcal{H}^d(\mathcal{C}') \\
&= \frac{\omega_{d-1}}{d} \ell^{d-1}(a+b) \\
&= \frac{\omega_{d-1}}{d} \ell^{d-1} h. \qquad (16)
\end{aligned}$$

Furthermore, since $a + b = h$, Pythagoras theorem gives

$$(r-b)^2 + \ell^2 = r^2 \quad \text{and} \quad (r'-a)^2 + \ell^2 = r'^2,$$

leading to

$$a = \frac{2rh - h^2}{2(r + r' - h)} = \frac{rh}{r+r'} + \frac{h^2}{r+r'-h}\left(\frac{r}{r+r'} - \frac{1}{2}\right).$$

Recalling that $r' \le r$, we may write

$$\frac{rh}{r+r'} \le a \le \frac{rh}{r+r'-h}.$$

Finally, since $\ell^2 = 2r'a - a^2$, we hence obtain

$$\ell^2 \geq a\left(2r' - \frac{rh}{r+r'-h}\right) \geq \frac{2r'rh}{r+r'}\left(1 - \frac{rh}{2r'(r+r'-h)}\right) \geq \frac{r'rh}{r+r'}.$$

Combining the equation above with (16) concludes the proof. □

Lemma A.10 (Mass of Intersection of Curved Balls)**.** *Let $x \in M$, and $T \in \mathbb{G}^{D,d}$. Let $O \in T$, and $r, R \geq 0$ be such that $\mathrm{B}_T(O,r) \subset \pi_T(\mathrm{B}(x,R) \cap M - x)$. For $A \geq C'_d r^{\frac{1-d}{2}}$, write*

$$h = \left(\frac{C_d f_{max}^4 \log n}{f_{\min}^5 (n-1)}\right)^{\frac{1}{d}}, \text{ and } \varepsilon_2 = \left(A\frac{f_{max}^4 \log n}{f_{\min}^5 (n-1)}\right)^{\frac{2}{d+1}}.$$

Then for n large enough, for all $\rho \geq r$ and $\Omega \in T$ such that $\|\Omega - O\| \leq r + \rho - \varepsilon_2$,

$$\int_{M \cap (\mathrm{B}(x,R) \setminus \mathrm{B}(x,h))} \mathbb{1}_{\pi_T(u-x) \in \mathrm{B}(O,r) \cap \mathrm{B}(\Omega,\rho)} f(u) \mathcal{H}^d(\mathrm{d}u) \geq A r^{\frac{d-1}{2}} C''_d \frac{f_{max}^4 \log n}{f_{\min}^4 (n-1)}.$$

Proof of Lemma A.10. As $\|\pi_T^t \circ \pi_T\|_{op} = \|\pi_T\|_{\mathrm{op}} \leq 1$, we have $\sqrt{|\det(\pi_T^t \circ \pi_T)|)} \leq 1$, so that the co-area formula [27, Theorem 3.1] entails that

$$\begin{aligned}
&\int_{M \cap (\mathrm{B}(x,R) \setminus \mathrm{B}(x,h))} \mathbb{1}_{\pi_T(u-x) \in \mathrm{B}(O,r) \cap \mathrm{B}(\Omega,\rho)} f(u) \mathcal{H}^d(\mathrm{d}u) \\
&\qquad \geq f_{\min} \int_{\pi_T(M \cap \mathrm{B}(x,R) - x)} \mathbb{1}_{\pi_T^{-1}(v) \notin \mathrm{B}(0,h)} \mathbb{1}_{\mathrm{B}(O,r) \cap \mathrm{B}(\Omega,\rho)}(v) \mathrm{d}v.
\end{aligned}$$

Since $\mathbb{1}_{\pi_T^{-1}(v) \notin \mathrm{B}(0,h)} \geq \mathbb{1}_{v \notin \mathrm{B}(0,h)}$, we get, provided A is large enough,

$$\begin{aligned}
&\int_{M \cap (\mathrm{B}(x,R) \setminus \mathrm{B}(x,h))} \mathbb{1}_{\pi_T(u-x) \in \mathrm{B}(O,r) \cap \mathrm{B}(\Omega,\rho)} f(u) \mathcal{H}^d(\mathrm{d}u) \\
&\quad \geq f_{\min} \int_{\pi_T(M \cap \mathrm{B}(x,R) - x)} \mathbb{1}_{v \notin \mathrm{B}(0,h)} \mathbb{1}_{\mathrm{B}(O,r) \cap \mathrm{B}(\Omega,\rho)}(v) \mathrm{d}v \\
&\quad \geq f_{\min} \int_{\mathrm{B}_T(0,r)} \mathbb{1}_{\mathrm{B}(O,r) \cap \mathrm{B}(\Omega,\rho)}(v) \mathrm{d}v - f_{\min} \omega_d h^d \\
&\quad \geq f_{\min} \left(\tilde{C}_d r^{\frac{d-1}{2}} A \frac{f_{\max}^4 \log n}{f_{\min}^5 (n-1)} - C_d \omega_d \frac{f_{\max}^4 \log n}{f_{\min}^5 (n-1)}\right) \\
&\quad \geq A r^{\frac{d-1}{2}} \tilde{C}'_d \frac{f_{\max}^4 \log n}{f_{\min}^4 (n-1)},
\end{aligned}$$

where the second to last inequality comes from Lemma B.12. □

C Proof of Lemma 5.3

Lemma 5.3. *Under the assumptions of Theorem 5.1, if $X_j \in J_{R_0,r,\rho}(X_i)$, then $\partial M \neq \emptyset$ and*

$$\mathrm{d}(X_j, \partial M) \leq 2r.$$

Proof of Lemma 5.3. Suppose that X_i is detected in the tangent space T_j. Then $\|X_i - X_j\| \leq r$, and there exists $\Omega \in T_j$ such that $\big\|\Omega - \pi_{T_j}(X_i - X_j)\big\| \geq \rho \geq \rho_-$ and $\mathcal{Y}_j \cap \mathrm{B}(\Omega, \big\|\Omega - \pi_{T_j}(X_i - X_j)\big\|) = \emptyset$. Since $\big\|\pi_{T_j}(X_i - X_j)\big\| \leq r$, it follows that $\|\Omega\| \geq \rho_- - r \geq 2r > r + \varepsilon_1$. Hence, define $u := \Omega / \|\Omega\|$, and $x := X_j + (r + \varepsilon_1)u$. As $\|\Omega - (x - X_j)\| = \|\Omega\| - r - \varepsilon_1 \leq \big\|\Omega - \pi_{T_j}(X_i - X_j)\big\| - \varepsilon_1$, we get

$$(\mathrm{B}(x, \varepsilon_1) - X_j) \cap \mathcal{Y}_j \subset \mathrm{B}\big(\Omega, \big\|\Omega - \pi_{T_j}(X_i - X_j)\big\|\big) \cap \mathcal{Y}_j = \emptyset. \tag{17}$$

From Equation (17), we now deduce that $x - X_j \notin \pi_{T_j}(\mathrm{B}(0, R_0) \cap (M - X_j))$. Indeed, if that was not the case, there would exist $y \in M \cap \mathrm{B}(X_j, R_0)$ such that $\pi_{T_j}(y - X_j) = x - X_j$. As $\mathrm{d_H}(M, \mathcal{X}_n) \leq \varepsilon_1$, there exists $X_k \in \mathrm{B}(y, \varepsilon_1) \cap \mathcal{X}_n$. Since $\|y - X_j\| \leq R_0 \leq \tau_{\min}/32$ and $\theta \leq 1/24$, Proposition A.4 yields that

$$\|X_k - X_j\| \leq \varepsilon_1 + \|y - X_j\| \leq \varepsilon_1 + \frac{\|x - X_j\|}{1 - \theta - \frac{\|y - X_j\|}{2\tau_{\min}}} \leq \varepsilon_1 + \frac{192}{181}(r + \varepsilon_1) \leq R_0,$$

and thus $X_k \in \mathrm{B}(X_j, R_0)$. By definition of $\mathcal{Y}_j$, this leads to $\pi_{T_j}(X_k - X_j) \in \mathcal{Y}_j$, and since

$$\big\|\pi_{T_j}(X_k - X_j) - (x - X_j)\big\| = \big\|\pi_{T_j}(X_k - y)\big\| \leq \|X_k - y\| \leq \varepsilon_1,$$

we get $\pi_{T_j}(X_k - X_j) \in \mathcal{Y}_j \cap (\mathrm{B}(x, \varepsilon_1) - X_j)$, which contradicts Equation (17). As a result, $x - X_j \notin \pi_{T_j}(\mathrm{B}(0, R_0) \cap (M - X_j))$, so that Lemma A.6 asserts that

$$\frac{4}{5} \min\{R_0 - 2\varepsilon_1, \mathrm{d}(X_j, \partial M)\} < \|x - X_j\| = r + \varepsilon_1.$$

As $4(R_0 - 2\varepsilon_1)/5 \geq R_0/4 \geq r + \varepsilon_1$ by assumption, the above inequality yields that $\mathrm{d}(X_j, \partial M) \leq 5(r + \varepsilon_1)/4 \leq 2r < \infty$, and in particular that $\partial M \neq \emptyset$, hence the result. □

D Tangent space estimation

D.1 Tangent space of the manifold

Proposition 3.2 (Tangent Space Estimation)**.** *Let $h = \big(C_d \frac{f_{\max}^4}{f_{\min}^5} \frac{\log n}{n-1}\big)^{\frac{1}{d}}$, for a large enough constant C_d. For n large enough so that $h \leq \frac{\tau_{\min}}{32} \wedge \frac{\tau_{\partial,\min}}{3} \wedge \frac{\tau_{\min}}{\sqrt{d}}$, with probability larger than $1 - 2\big(\frac{1}{n}\big)^{\frac{2}{d}}$, we have*

$$\max_{1 \leq i \leq n} \angle(T_{X_i} M, \hat{T}_i) \leq C_d \frac{f_{\max}}{f_{\min}} \frac{h}{\tau_{\min}}.$$

Proof of Proposition 3.2. We let $h = \Big(\frac{\kappa}{f_{\min}} \frac{\log n}{n-1}\Big)^{\frac{1}{d}}$, where $\kappa > 1$ will be fixed later, and assume that n is large enough so that $h \leq \frac{\tau_M}{32} \wedge \frac{\tau_{\partial M}}{3} \wedge \frac{\tau_M}{\sqrt{d}}$. Without loss of generality we consider the case where $i = 1$ and $X_1 = 0$. We let $x \in \mathrm{B}(0, h) \cap M$ be such that $\mathrm{B}(x, h/4) \cap \partial M = \emptyset$, according to Corollary B.6. Slightly differing from the notation in Proposition A.4, for any vector $u \in \mathbb{R}^d$, we denote by $u_T = \pi_{T_x M}(u)$ and $u_\perp = \pi_{(T_x M)^\perp}(u)$. For short, we also write $p(h) := P(\mathrm{B}(0, h))$ and $p_n(h) := P_n(\mathrm{B}(0, h))$, where $P_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ stands for the empirical measure. The proof of Proposition 3.2 will make use of the following concentration result, borrowed from [2].

Lemma D.1 ([2, Lemma 9.5])**.** *Write*

$$\Sigma(h) := \mathbb{E}\left(X_T (X_T)^t \mathbb{1}_{\mathrm{B}(0,h)}(X)\right).$$

Then for n large enough, with probability larger than $1-2\left(\frac{1}{n}\right)^{1+\frac{2}{d}}$, we have,

$$p_n(h) \leq 2p(h) + \frac{10(2+\frac{2}{d})\log n}{n-1},$$

and

$$\left\| \frac{1}{n-1}\sum_{i=2}^{n}(X_i)_T(X_i)_T^t \mathbb{1}_{\mathrm{B}(0,h)}(X_i) - \Sigma(h) \right\|_F \leq C_d \frac{f_{\max}}{f_{\min}\sqrt{\kappa}} p(h)h^2.$$

We now assume that the event described by Lemma D.1 occurs. We may decompose the local covariance matrix as

$$\frac{1}{n-1}\sum_{i=2}^{n}(X_i)(X_i)^t = \sum_{i=2}^{n}(X_i)_T(X_i)_T^t + R_1,$$

where

$$R_1 := \frac{1}{n-1}\sum_{i=2}^{n}\left[(X_i)_T(X_i)_\perp^t + (X_i)_\perp(X_i)_T^t + (X_i)_\perp(X_i)_\perp^t\right].$$

Since $\mathrm{B}(0,h) \subset \mathrm{B}(x,2h)$, we have $\|(X_i)_T\| \leq h$ and, according to [27, Theorem 4.18], $\|(X_i)_\perp\| \leq \|(X_i-x)_\perp\| + \|(x-0)_\perp\| \leq \frac{3h^2}{\tau_M}$. Thus, $\|R_1\|_F \leq \frac{9h^3}{\tau_M}p_n(h) \leq C_d\frac{f_{\max}h^{d+3}}{\tau_M}$, according to Lemma D.1.

Next, using Lemma D.1 again, we have

$$\lambda_{\min}\left(\frac{1}{n-1}\sum_{i=2}^{n}(X_i)_T(X_i)_T^t\mathbb{1}_{\mathrm{B}(0,h)}(X_i)\right) \geq \lambda_{\min}\left(\Sigma(h)\right) - C_d\frac{f_{\max}^2}{f_{\min}\sqrt{\kappa}}h^{d+2}.$$

On the other hand, for $u \in T_xM$, we have

$$\begin{aligned}
u^t\Sigma(h)u &= \int_{\mathrm{B}(0,h)\cap M}\langle u, y_T\rangle^2 f(y)\mathcal{H}^d(\mathrm{d}y) \\
&\geq f_{\min}\int_{\mathrm{B}(x,h/4)\cap M}\langle u, y_T\rangle^2 f(y)\mathcal{H}^d(\mathrm{d}y) \\
&\geq f_{\min}\int_{\mathrm{B}_d(0,h/4)}\langle u, \exp_x(v)_T - x_T + x_T\rangle^2 \left|\det\left(d_v(\exp_x)\right)\right|\mathrm{d}v,
\end{aligned}$$

according to [2, Propositions 8.5 and 8.6]. Moreover, [2, Proposition 8.7] ensures that we have $|\det(d_v(\exp_x))| \geq c_d$ provided that $\|v\| \leq h/4 \leq \tau_M/4$, and [2, Proposition 8.6] gives $\exp_x(v) = x + v + R(v)$, with $\|R(v)\| \leq \frac{5\|v\|^2}{8\tau_M}$, under the same condition. Thus,

$$\begin{aligned}
u^t\Sigma(h)u &\geq c_d f_{\min}\int_{\mathrm{B}_d(0,h/4)}\langle u, v + R(v)_T + x_T\rangle^2\,\mathrm{d}v \\
&\geq \frac{1}{2}c_d f_{\min}\int_{\mathrm{B}_d(0,h/4)}\langle u, v + x_T\rangle^2\,\mathrm{d}v \\
&\qquad - 3c_d f_{\min}\int_{\mathrm{B}_d(0,h/4)}\left(\frac{5\|v\|^2}{8\tau_{\min}}\right)^2\mathrm{d}v.
\end{aligned}$$

Denoting by σ_{d-1} the surface of the $(d-1)$-dimensional unit sphere and using polar coordinates yields

$$\int_{\mathrm{B}_d(0,h/4)}\langle u, v + x_T\rangle^2\,\mathrm{d}v \geq \int_{\mathrm{B}_d(0,h/4)}\langle u, v\rangle^2\,\mathrm{d}v \geq \left(\frac{h}{4}\right)^{d+2}\frac{1}{d(d+2)}\sigma_{d-1},$$

and

$$\int_{\mathrm{B}_d(0,h/4)} \left(\frac{5\|v\|^2}{8\tau_{\min}}\right)^2 \mathrm{d}v \leq \left(\frac{5}{8}\right)^2 \frac{\sigma_{d-1}}{(d+4)\tau_M^2} \left(\frac{h}{4}\right)^{d+4}.$$

Since $h \leq \tau_M/\sqrt{d}$, it follows that

$$\lambda_{\min}(\Sigma(h)) \geq c_d f_{\min} h^{d+2},$$

for some positive constant c_d. Gathering all pieces and using [2, Theorem 10.1] leads to

$$\angle(T_x M, \hat{T}_i) \leq C_d \frac{f_{\max} h}{\tau_M(c_d f_{\min} - C_d(f_{\max}^2/(f_{\min}\sqrt{\kappa})))}.$$

Thus, choosing $\kappa = C_d \left(\frac{f_{\max}}{f_{\min}}\right)^4$, for C_d large enough, gives

$$\angle(T_x M, \hat{T}_i) \leq C_d \frac{f_{\max} h}{f_{\min}\tau_M}.$$

Noting that $\angle(T_0 M, T_x M) \leq 2h/\tau_M$ from Proposition A.2 and lemma A.1, the result of Proposition 3.2 follows after using a union bound. □

D.2 Tangent space of the boundary

Corollary 3.11 (Boundary's Tangent Space Estimation)**.** *Under the assumptions of Proposition 3.2 and Theorem 3.7 we have, for n large enough, with probability larger than $1 - 4n^{-\frac{2}{d}}$,*

$$\max_{X_i \in \mathcal{Y}_{R_0,r,\rho}} \angle(T_{\pi_{\partial M}(X_i)}\partial M, \hat{T}_{\partial,i}) \leq \frac{20r}{\sqrt{(\tau_{\min} \wedge \tau_{\partial,\min})R_0}}.$$

Thus, choosing $R_0 = \frac{\tau_{\min}\wedge\tau_{\partial,\min}}{40}$ and $r = r_-$ yields

$$\max_{X_i \in \mathcal{Y}_{R_0,r_-,\rho}} \angle(T_{\pi_{\partial M}(X_i)}\partial M, \hat{T}_{\partial,i}) \leq \left(C_d \frac{f_{\max}^5}{f_{\min}^5} \frac{\log n}{n f_{\min}(\tau_{\min} \wedge \tau_{\partial,\min})^d}\right)^{\frac{1}{d+1}}.$$

Proof of Corollary 3.11. Under the assumptions of Theorem 3.7 and Proposition 3.2, we let $X_i \in \mathcal{Y}_{R_0,r,\rho}$, $\varepsilon_{\partial M} = \left(C_d R_0 \frac{f_{\max}}{f_{\min}^2}\frac{\log n}{n}\right)^{\frac{1}{d+1}}$, and $h = \left(C_d \frac{f_{\max}^4}{f_{\min}^5}\frac{\log n}{n-1}\right)^{\frac{1}{d}}$ so that with probability larger than $1 - 4n^{-2/d}$, we have

$$\angle(\eta_{\pi_{\partial M}(X_i)}, \tilde{\eta}_i) \leq \frac{\varepsilon_{\partial M}}{R_0} \quad \text{and} \quad \angle(T_{X_i} M, \hat{T}_i) \leq C_d \frac{f_{\max}}{f_{\min}} \frac{h}{\tau_{\min}}.$$

Combining Theorem 3.7 (i) with Lemma A.1 and Proposition A.2 entails

$$\begin{aligned}\angle(T_{\pi_{\partial M}(X_i)} M, \hat{T}_i) &\leq \angle(T_{\pi_{\partial M}(X_i)} M, T_{X_i} M) + \angle(T_{X_i} M, \hat{T}_i) \\ &\leq 2\frac{\varepsilon_{\partial M}^2}{R_0} + C_d \frac{f_{\max}}{f_{\min}} \frac{h}{\tau_{\min}} \leq C_d \varepsilon_{\partial M},\end{aligned}$$

for n large enough. Finally, since

$$\angle(T_{\pi_{\partial M}(X_i)}\partial M, \hat{T}_{\partial,i}) \leq \angle(T_{\pi_{\partial M}(X_i)} M, \hat{T}_i) + \angle(\eta_{\pi_{\partial M}(X_i)}, \tilde{\eta}_i),$$

the bound follows. □

E Local linear patches

Theorem 5.6 (Estimation with Local Linear Patches)**.** *Write* $r_0 := (\tau_{\min} \wedge \tau_{\partial,\min})/40$*, let* $\varepsilon_0, a, \delta \geq 0$*, and* $0 \leq \theta, \theta' \leq 1/16$*. Assume that we have:*

1. *A point cloud* $\mathcal{X}_n \subset M$ *such that* $\mathrm{d_H}(M, \mathcal{X}_n) \leq \varepsilon_0$*,*
2. *Estimated tangent spaces* $(T_i)_{1\leq i\leq n}$ *such that* $\max_{1\leq i\leq n} \angle(T_{X_i}M, T_i) \leq \theta$*,*
3. *A subset of boundary observations* $\mathcal{X}_\partial \subset \mathcal{X}_n$ *such that*

$$\max_{x\in\partial M} \mathrm{d}(x, \mathcal{X}_\partial) \leq \delta \text{ and } \max_{x\in\mathcal{X}_\partial} \mathrm{d}(x, \partial M) \leq a\delta^2,$$

from which we build interior observations

$$\mathring{\mathcal{X}}_{\varepsilon_{\partial M}} := \{X_i \in \mathcal{X}_n \mid \mathrm{d}(X_i, \mathcal{X}_\partial) \geq \varepsilon_{\partial M}/2\}.$$

4. *Estimated unit normal vectors* $(\eta_i)_{1\leq i\leq n}$ *on* $\mathcal{X}_\partial$ *such that* $\max_{X_i\in\mathcal{X}_\partial} \|\eta_i - \eta_{\pi_{\partial M}(X_i)}\| \leq \theta'$*.*

Let $\mathbb{M} = \mathbb{M}(\mathcal{X}_n, \mathcal{X}_\partial, T, \eta)$ *be defined as* $\mathbb{M} := \mathbb{M}_{\mathrm{Int}} \cup \mathbb{M}_\partial$*, with*

$$\mathbb{M}_{\mathrm{Int}} := \bigcup_{X_i\in\mathring{\mathcal{X}}_{\varepsilon_{\partial M}}} X_i + \mathrm{B}_{T_i}(0, \varepsilon_{\mathring{M}}),$$
$$\mathbb{M}_\partial := \bigcup_{X_i\in\mathcal{X}_\partial} (X_i + \mathrm{B}_{T_i}(0, \varepsilon_{\partial M})) \cap \{z, \langle z - X_i, \eta_i\rangle \leq 0\},$$

Then if $\varepsilon_{\partial M} \leq r_0/2$*,* $\varepsilon_0 \leq \varepsilon_{\mathring{M}} \leq \varepsilon_{\partial M}/6$ *, and* $\max\{\delta, a\delta^2\} \leq \varepsilon_{\partial M}/6$*, we have*

$$\mathrm{d_H}\big(M, \mathbb{M}\big) \leq \begin{cases} \varepsilon_{\mathring{M}}\left(\theta + \varepsilon_{\mathring{M}}/\tau_{\min}\right) & \text{if } \partial M = \emptyset, \\ 2a\delta^2 + 8\varepsilon_{\partial M}\left(\theta + \theta' + \varepsilon_{\partial M}/r_0\right) & \text{if } \partial M \neq \emptyset. \end{cases}$$

Proof of Theorem 5.6. First, note that the choice $r_0 = (\tau_{\min} \wedge \tau_{\partial,\min})/40$ satisfies the requirements of Lemma A.7, for a radius $R_0 = \tau_{\min}/16$. For short, let $\mathbb{M} := \mathbb{M}(\mathcal{X}_n, \mathcal{X}_\partial, T, \eta)$.

- Let $x \in M$ be fixed. We bound $\mathrm{d}(x, \mathbb{M})$ depending on its closeness to ∂M.
  - First assume that $\mathrm{d}(x, \partial M) \leq \varepsilon_{\partial M} - \delta$. Then $\mathrm{d}(x, \mathcal{X}_\partial) \leq \varepsilon_{\partial M}$, and we let $X_{i_0} \in \mathcal{X}_\partial$ be such that $\|x - X_{i_0}\| \leq \varepsilon_{\partial M}$. Without loss of generality we may assume that $i_0 = 1$. Let $\mathbb{P}_1 := X_1 + \mathrm{B}_{T_1}(0, \varepsilon_{\partial M}) \cap \{z, \langle z - X_1, \eta_1\rangle \leq 0\} \subset \mathbb{M}_\partial$ denote the half-patch at X_1.
    From Proposition A.4, we have

$$\|\pi_{X_1+T_1}(x) - x\| \leq \varepsilon_{\partial M}\left(\theta + \frac{\varepsilon_{\partial M}}{2\tau_{\min}}\right). \tag{18}$$

    As a result, if $\pi_{X_1+T_1}(x) \in \mathbb{P}_1$, then $\mathrm{d}(x, \mathbb{M}) \leq \|\pi_{X_1+T_1}(x) - x\|$ yields the desired bound. Otherwise, if $\pi_{X_1+T_1}(x) \notin \mathbb{P}_1$, since $\mathrm{d}(x, \mathbb{P}_1) \leq \|x - X_1\| \leq \varepsilon_{\partial M}$, we can decompose $\pi_{X_1+T_1}(x)$ as $\pi_{X_1+T_1}(x) = X_1 + \alpha\eta_1 + \beta v$, with unit $v \in T_1 \cap \mathrm{span}(\eta_1)^\perp$, and $\alpha = \mathrm{d}(\pi_{X_1+T_1}(x), \mathbb{P}_1) > 0$ such that $\alpha^2 + \beta^2 \leq \varepsilon_{\partial M}^2$. Writing $x_1 := \pi_{\partial M}(X_1)$, triangle inequality ensures that

$$\|x - x_1\| \leq \|x - X_1\| + \|X_1 - x_1\| \leq \varepsilon_{\partial M} + a\delta^2 \leq \tau_{\min}/32.$$

From Lemma A.1 and proposition A.2, we also have

$$\angle(T_{x_1}M, T_1) \leq \angle(T_{x_1}M, T_{X_1}M) + \angle(T_{X_1}M, T_1) \leq 2\varepsilon_{\partial M}/\tau_{\min} + \theta \leq 1/8.$$

As a result, Lemma A.7 applies and gives

$$\pi_{T_1}(x - X_1) \in \mathrm{B}_{T_1}(0, \varepsilon_{\partial M}) \cap (\mathrm{B}_{T_1}(\pi_{T_1}(x_1 - X_1) + r_0\eta_{x_1}, r_0))^c.$$

Thus, we have

$$\begin{aligned} r_0 &\leq \|\alpha\eta_1 + \beta v - r_0\eta_{x_1} + \pi_{T_1}(X_1 - x_1)\| \\ &\leq \|\alpha\eta_1 + \beta v - r_0\eta_{x_1}\| + a\delta^2, \end{aligned}$$

which, since $a\delta^2 \leq r_0$ and $\alpha^2 + \beta^2 \leq \varepsilon_{\partial M}$, leads to

$$\begin{aligned} (r_0 - a\delta^2)^2 &\leq \|(\alpha\eta_1 + \beta v) - r_0\eta_{x_1}\|^2 \\ &\leq \varepsilon_{\partial M}^2 + r_0^2 - 2r_0\alpha\langle\eta_1, \eta_{x_1}\rangle - 2r_0\beta\langle v, \eta_{x_1}\rangle. \end{aligned}$$

As $\langle\eta_1, \eta_{x_1}\rangle = 1 - \|\eta_1 - \eta_{x_1}\|^2/2 \geq 1 - \theta'^2/2 > 0$ and $|\langle v, \eta_{x_1}\rangle| = |\langle v, \eta_1 - \eta_{x_1}\rangle| \leq \theta'$, we deduce that

$$\alpha = \mathrm{d}(\pi_{X_1+T_1}(x), \mathbb{P}_1) \leq \frac{\varepsilon_{\partial M}^2 + 2r_0a\delta^2 + r_0\varepsilon_{\partial M}\theta'}{2r_0(1 - \theta'^2/2)} \leq \frac{\varepsilon_{\partial M}^2}{r_0} + 2a\delta^2 + \varepsilon_{\partial M}\theta'.$$

At the end of the day, combining the above inequality with (18) yields the bound

$$\mathrm{d}(x, \mathbb{M}) \leq 2a\delta^2 + \varepsilon_{\partial M}\left(\theta + \theta' + \frac{2\varepsilon_{\partial M}}{r_0}\right), \tag{19}$$

which also holds if $\pi_{X_1+T_1}(x) \in \mathbb{P}_1$.

– Now, assume that $\mathrm{d}(x, \partial M) > \varepsilon_{\partial M} - \delta$. Let X_{i_0} denote the closest point to x in $\mathcal{X}_n$, with $i_0 = 1$ without loss of generality. Since $\|x - X_1\| \leq \varepsilon_0$, we deduce that

$$\begin{aligned} \mathrm{d}(X_1, \mathcal{X}_\partial) &\geq \mathrm{d}(x, \mathcal{X}_\partial) - \|x - X_1\| \\ &\geq \mathrm{d}(x, \partial M) - a\delta^2 - \varepsilon_0 \\ &\geq \varepsilon_{\partial M} - \delta - \varepsilon_0 - a\delta^2 \\ &\geq \varepsilon_{\partial M}/2. \end{aligned}$$

Thus $X_1 \in \mathring{\mathcal{X}}_{\varepsilon_{\partial M}}$, and therefore $\mathbb{P}_1 := X_1 + \mathrm{B}_{T_1}(0, \varepsilon_{\mathring{M}})$ is a patch of $\mathbb{M}_{\mathrm{Int}} \subset \mathbb{M}$. Because $\varepsilon_{\mathring{M}} \geq \varepsilon_0$, the point $\pi_{X_1+T_1}(x)$ belongs to $\mathbb{P}_1$, so that $\mathrm{d}(x, \mathbb{M}) \leq \|\pi_{X_1+T_1}(x) - x\|$. Using Proposition A.4 again, we get

$$\mathrm{d}(x, \mathbb{M}) \leq \varepsilon_0\left(\theta + \frac{\varepsilon_0}{2\tau_{\min}}\right). \tag{20}$$

- Let now $x \in \mathbb{M}$ be fixed. We bound $\mathrm{d}(x, M)$ depending on whether x belongs to a "boundary patch" (i.e. to $\mathbb{M}_\partial$) or an "interior patch" (i.e. to $\mathbb{M}_{\mathrm{Int}}$).

  – Assume that $x \in \mathbb{M}_\partial$ belongs to "boundary patch". That is, without loss of generality, $x \in X_1 + \mathrm{B}_{T_1}(0, \varepsilon_{\partial M}) \cap \{z, \langle z - X_1, \eta_1\rangle \leq 0\}$ with $X_1 \in \mathcal{X}_\partial$. Define $x_1 := \pi_{\partial M}(X_1)$,

$$x_1^* := \pi_{\pi_{X_1+T_1}(\partial M \cap \mathrm{B}(X_1, \tau_{\min}/16))}(X_1),$$

and let $x_1' \in \partial M \cap \mathrm{B}(X_1, \tau_{\min}/16)$ be such that $\pi_{X_1+T_1}(x_1') = x_1^*$. According to Corollary A.8, we have $x_1^* - X_1 = \|x_1^* - X_1\|\,\eta_1^*$, where η_1^* is the unit vector of $Nor(x_1', M) \cap T_1$. Furthermore, Proposition A.2, Proposition B.1 and Proposition B.8 combined yield the bound

$$\|\eta_1^* - \eta_{x_1'}\| \leq \sqrt{2}\angle(T_{x_1'}M, T_1) \leq \sqrt{2}(\theta + 2\|x_1' - X_1\|/\tau_{\min}).$$

Furthermore, by definition of x_1^* and the fact that $x_1 \in \partial M \cap \mathrm{B}(X_1, \tau_{\min}/16)$, we also have

$$\|X_1 - x_1^*\| \leq \|\pi_{T_1}(X_1 - x_1)\| \leq \|X_1 - x_1\| = \mathrm{d}(X_1, \partial M) \leq a\delta^2.$$

As $\|X_1 - x_1'\| \leq \tau_{\min}/16$, Proposition A.4 ensures that $\|X_1 - x_1^*\| \geq \|X_1 - x_1'\|\,(1 - \theta - 1/32)$, which leads to $\|X_1 - x_1'\| \leq 2a\delta^2$ and hence to $\|x_1 - x_1'\| \leq 3a\delta^2 \leq (\tau_{\min} \wedge \tau_{\partial,\min})/32$. As a result, Proposition A.3 applies and asserts that

$$\|\eta_{x_1} - \eta_{x_1'}\| \leq \frac{9\|x_1 - x_1'\|}{\tau_{\min} \wedge \tau_{\partial,\min}} \leq \frac{27a\delta^2}{\tau_{\min} \wedge \tau_{\partial,\min}}.$$

Gathering all the pieces together, we obtain

$$\begin{aligned}
\|\eta_1^* - \eta_1\| &\leq \|\eta_1^* - \eta_{x_1'}\| + \|\eta_{x_1'} - \eta_{x_1}\| + \|\eta_{x_1} - \eta_1\| \\
&\leq \sqrt{2}\theta + \theta' + \frac{(27 + 4\sqrt{2})a\delta^2}{\tau_{\min} \wedge \tau_{\partial,\min}} \\
&\leq \sqrt{2}\theta + \theta' + \frac{a\delta^2}{r_0} \\
&:= \theta''.
\end{aligned}$$

Now, if $x \in \mathrm{B}_{X_1+T_1}(x_1^* - r_0\eta_1^*, r_0)$, we have $\mathrm{d}(x, \mathrm{B}_{X_1+T_1}(x_1^* - r_0\eta_1^*, r_0)) = 0$. Otherwise, if $x \notin \mathrm{B}_{X_1+T_1}(x_1^* - r_0\eta_1^*, r_0)$, we have

$$\mathrm{d}(x, \mathrm{B}_{X_1+T_1}(x_1^* - r_0\eta_1^*, r_0)) = \|x - (x_1^* - r_0\eta_1^*)\| - r_0 > 0.$$

We may hence write

$$\begin{cases} x - X_1 = -\alpha\eta_1 + \beta v \text{ with } \alpha \geq 0, \alpha^2 + \beta^2 \leq \varepsilon_{\partial M}^2, \text{ and unit } v \in T_1 \cap \mathrm{span}(\eta_1)^\perp, \\ x_1^* - X_1 = t\eta_1^* \text{ with } 0 \leq t \leq a\delta^2 \text{ and } \|\eta_1 - \eta_1^*\| \leq \theta''. \end{cases}$$

Since $\langle \eta_1, \eta_1^* \rangle \geq 0$ and $|\langle v, \eta_1^* \rangle| = |\langle v, \eta_1^* - \eta_1 \rangle| \leq \theta''$, it follows that

$$\begin{aligned}
\|x - (x_1^* - r_0\eta_1^*)\|^2 &= \|(x - X_1) + (r_0 - t)\eta_1^*\|^2 \\
&\leq \varepsilon_{\partial M}^2 + 2(r_0 - t)(\langle -\alpha\eta_1, \eta_1^* \rangle + \langle \beta v, \eta_1^* \rangle) + (r_0 - t)^2 \\
&\leq \varepsilon_{\partial M}^2 + 2\varepsilon_{\partial M}(r_0 - t)\theta'' + (r_0 - t)^2.
\end{aligned}$$

Therefore, no matter whether or not x belongs to $\mathrm{B}_{X_1+T_1}(x_1^* - r_0\eta_1^*, r_0)$, we have

$$\begin{aligned}
\mathrm{d}(x, \mathrm{B}_{X_1+T_1}(x_1^* - r_0\eta_1^*, r_0)) &\leq \varepsilon_{\partial M}\theta'' + \frac{\varepsilon_{\partial M}^2}{2(r_0 - a\delta^2)} \\
&\leq \varepsilon_{\partial M}\theta'' + \frac{\varepsilon_{\partial M}^2}{r_0}.
\end{aligned}$$

From the left-hand side inclusion of Lemma A.7, we hence get the existence of some $y \in \mathrm{B}(x_1', \tau_{\min}/16) \cap M$ such that

$$\|x - \pi_{X_1+T_1}(y)\| \leq \varepsilon_{\partial M}\theta'' + \frac{\varepsilon_{\partial M}^2}{r_0}.$$

We will now show that this point $y \in M$ is close to x.
For this, a first (rough) bound on $\|y - X_1\|$ may be derived, using $\|y - X_1\| \leq \|y - x_1'\| + \|x_1' - X_1\| \leq \tau_{\min}/16 + 2a\delta^2 \leq \tau_{\min}/8$. According to Proposition A.4, we have

$$\|y - X_1\| \leq \frac{\|\pi_{T_1}(y - X_1)\|}{1 - \theta - \|y - X_1\|/(2\tau_{\min})} \leq 2\|\pi_{T_1}(y - X_1)\|,$$

which, by using the other bound of Proposition A.4, leads to

$$\|y - \pi_{X_1+T_1}(y)\| \leq 2\|\pi_{T_1}(y - X_1)\| \left(\theta + \frac{\|\pi_{T_1}(y - X_1)\|}{\tau_{\min}}\right),$$

Hence, further bounding

$$\begin{aligned}\|\pi_{T_1}(y - X_1)\| &\leq \|x - X_1\| + \|x - \pi_{X_1+T_1}(y)\| \\ &\leq \varepsilon_{\partial M} + \varepsilon_{\partial M}\theta'' + \frac{\varepsilon_{\partial M}^2}{r_0} \\ &\leq 2\varepsilon_{\partial M}\end{aligned}$$

since $\theta'' \leq 1/2$ and $\varepsilon_{\partial M} \leq r_0/2$, we finally obtain

$$\begin{aligned}\|x - y\| &\leq \|x - \pi_{X_1+T_1}(y)\| + \|y - \pi_{X_1+T_1}(y)\| \\ &\leq \varepsilon_{\partial M}\theta'' + \frac{\varepsilon_{\partial M}^2}{r_0} + 4\varepsilon_{\partial M}\left(\theta + \frac{2\varepsilon_{\partial M}}{\tau_{\min}}\right) \\ &\leq 8\varepsilon_{\partial M}\left(\theta + \theta' + \frac{\varepsilon_{\partial M}}{r_0}\right),\end{aligned}$$

where we used that $a\delta^2 \leq \varepsilon_{\partial M}$. In particular, we have

$$\mathrm{d}(x, M) \leq 8\varepsilon_{\partial M}\left(\theta + \theta' + \frac{\varepsilon_{\partial M}}{r_0}\right). \tag{21}$$

– Assume that $x \in \mathbb{M}_{\mathrm{Int}}$ belongs to an "interior patch". That is, without loss of generality, $x \in X_1 + \mathrm{B}_{T_1}(0, \varepsilon_{\mathring{M}})$ with $\mathrm{d}(X_1, \mathcal{X}_\partial) \geq \varepsilon_{\partial M}/2$. We have $\mathrm{d}(X_1, \partial M) \geq \varepsilon_{\partial M}/2 - \delta \geq 3\varepsilon_{\mathring{M}}/2$, so that an applying Lemma A.6 at X_1 provides the existence of some $y \in M \cap \mathrm{B}(X_1, \varepsilon_{\mathring{M}})$ such that $x = \pi_{X_1+T_1}(y)$. Thus, Proposition A.4 entails

$$\mathrm{d}(x, M) \leq \|y - x\| = \|(y - X_1)^\perp\| \leq \varepsilon_{\mathring{M}}\left(\theta + \frac{\varepsilon_{\mathring{M}}}{2\tau_{\min}}\right). \tag{22}$$

To conclude the proof of Theorem 5.6, we combine the above results as follows.

(i) If $\partial M = \emptyset$, then $\mathrm{d}(x, \partial M) = \infty$ for all $x \in \mathbb{R}^D$, so that $\mathcal{X}_\partial = \emptyset$ and hence $\mathbb{M}_\partial = \emptyset$. As a result, $\mathrm{d_H}(M, \mathbb{M})$ is bounded by the maximum of Equations (20) and (22). The requirement $\varepsilon_0 \leq \varepsilon_{\mathring{M}}$ ensures that

$$\mathrm{d_H}(M, \mathbb{M}) \leq \varepsilon_{\mathring{M}}\left(\theta + \frac{\varepsilon_{\mathring{M}}}{2\tau_{\min}}\right).$$

(ii) If $\partial M \neq \emptyset$, then $\mathrm{d_H}(M, \mathbb{M})$ is bounded by the maximum of Equations (19) to (22). This boils down to

$$\mathrm{d_H}(M, \mathbb{M}) \leq 2a\delta^2 + 8\varepsilon_{\partial M}\left(\theta + \theta' + \frac{\varepsilon_{\partial M}}{r_0}\right).$$

□

F Proofs of the minimax lower bounds

The minimax lower bounds (Theorems 3.13 and 3.16) will be proven using the standard Bayesian arguments relying on hypotheses comparison method. This is usually referred to as Le Cam's method. It involves the total variation distance, for which we recall a definition.

Definition F.1 (Total Variation)**.** For any two Borel probability distributions P_0, P_1 over $\mathbb{R}^D$, the total variation between them is defined as

$$\mathrm{TV}(P_0, P_1) := \frac{1}{2}\int_{\mathbb{R}^D} |f_1 - f_0| d\mu,$$

where μ is a σ-finite measure dominating P_0 and P_1, with respective densities f_0 and f_1.

In the context of manifold and boundary estimation for the Hausdorff distance $\mathrm{d_H}$, Le Cam's lemma [49] writes as follows.

Lemma F.2. *Fix an integer $n \geq 1$ and write $\mathcal{P} = \mathcal{P}^{d,D}_{\tau_{\min}, \tau_{\partial,\min}}(f_{\min}, f_{\max})$.*

(i) Then for all $P_0, P_1 \in \mathcal{P}$ with respective supports M_0 and M_1,

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}\left[\mathrm{d_H}\big(M, \hat{M}\big)\right] \geq \frac{1}{2}\mathrm{d_H}(M_0, M_1)\left(1 - \mathrm{TV}(P_0, P_1)\right)^n,$$

where the infimum ranges among all the estimators $\hat{M} = \hat{M}(X_1, \ldots, X_n)$.

(ii) If in addition, ∂M_0 and ∂M_1 are non-empty,

$$\inf_{\hat{B}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}\left[\mathrm{d_H}\big(\partial M, \hat{B}\big)\mathbb{1}_{\partial M \neq \emptyset}\right] \geq \frac{1}{2}\mathrm{d_H}(\partial M_0, \partial M_1)\left(1 - \mathrm{TV}(P_0, P_1)\right)^n,$$

where the infimum ranges among all the estimators $\hat{B} = \hat{B}(X_1, \ldots, X_n)$.

Proof of Lemma F.2. Apply [49, Lemma 1] with loss function $\mathrm{d_H}$, model $\mathcal{P}$, parameters of interest $\theta(P) = \mathrm{Supp}(P)$ and $\theta(P) = \partial\big(\mathrm{Supp}(P)\big)$ respectively, and conclude with the bound $(1 - \mathrm{TV}(P_0^n, P_1^n)) \geq (1 - \mathrm{TV}(P_0, P_1))^n$. □

Aiming at applying Lemma F.2, we shall first describe how to construct hypotheses P_0 and P_1 that belong to the models, close in total variation distance but with supports (or boundary) far away in Hausdorff distance.

F.1 Hypotheses with empty boundary

To do so in the boundariless case $\tau_{\partial,\min} = \infty$, we will use a structural stability result of the family of model. We recall that $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm, that is $\|A\|_{\mathrm{op}} = \max_{\|v\|=1} \|Av\|$ for all $A \in \mathbb{R}^{D\times D}$.

Proposition F.3 (Reach Stability). *Let $M \in \mathcal{M}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}$ and $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ be a $\mathcal{C}^2$ map such that $\lim_{\|x\|\to\infty} \|\Phi(x)\| = \infty$. Assume that $\sup_{x\in\mathbb{R}^D} \|I_D - d_x\Phi\|_{\mathrm{op}} \leq 1/10$. Then Φ is a global diffeomorphism, and the image $\Phi(M)$ of M by Φ satisfies:*

- $\partial\Phi(M) = \Phi(\partial M)$,
- *If $\sup_{x\in\mathbb{R}^D} \left\|d^2_x\Phi\right\|_{\mathrm{op}} \leq 1/(2\tau_{\min})$, then $\tau_{\Phi(M)} \geq \tau_{\min}/2$,*
- *If $\sup_{x\in\mathbb{R}^D} \left\|d^2_x\Phi\right\|_{\mathrm{op}} \leq 1/(2\tau_{\partial\min})$, then $\tau_{\partial\Phi(M)} \geq \tau_{\partial,\min}/2$.*

The proof is to be found in Appendix G.1. Essentially, the class $\left\{\mathcal{M}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}\right\}_{\tau_{\min},\tau_{\partial,\min}}$ is stable up to $\mathcal{C}^2$-diffeomorphism, with explicit bounds on the parameters. From there, we consider P_0 over a boundariless manifold $M_0 \in \mathcal{M}^{d,D}_{2\tau_{\min},\infty}$, and P_1 over M_1 that is obtained by bumping M_1 locally (see Figure 10). The method is similar to that of [3, Lemma 5], with an explicit dependency in the parameters of the model.

Proposition F.4 (Hypotheses with Empty Boundary). *Assume that $f_{\min} \leq c_d/\tau^d_{\min}$ and $c'_d/\tau^d_{\min} \leq f_{\max}$, for some small enough $c_d, (c'_d)^{-1} > 0$.*

If $d \leq D-1$, then for all $n \geq C_d/(f_{\min}\tau^d_{\min})$, there exist $P_0, P_1 \in \mathcal{P}^{d,D}_{\tau_{\min},\infty}(f_{\min}, f_{\max})$ with boundariless supports M_0 and M_1 such that

$$\mathrm{TV}(P_0, P_1) \leq \frac{1}{n} \quad \text{and} \quad \mathrm{d_H}(M_0, M_1) \geq C'_d\tau_{\min}\left(\frac{1}{f_{\min}\tau^d_{\min}n}\right)^{2/d}.$$

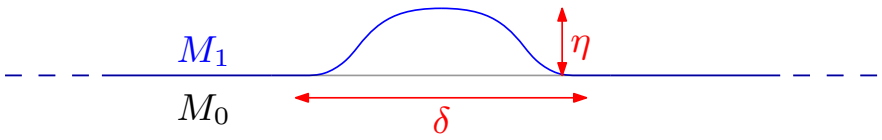


Figure 10: Boundariless supports M_0 and M_1 of Proposition F.4 for $d = 1$ and $D = 2$. Here, the total variation between the associated uniform distributions is of order $\mathrm{TV}(P_0, P_1) \asymp f_{\min}\mathcal{H}^d(M_0 \triangle M_1) \asymp f_{\min}\delta^d$ and Hausdorff distance $\mathrm{d_H}(M_0, M_1) = \eta$. The reach bound forces the bump to have height $\eta \lesssim \delta^2/\tau_{\min}$, so that optimal parameter choices yield:

$$\delta \asymp \left(\frac{1}{f_{\min}n}\right)^{1/d} \quad \text{and} \quad \eta \asymp \frac{\delta^2}{\tau_{\min}} \asymp \tau_{\min}\left(\frac{1}{f_{\min}\tau^d_{\min}n}\right)^{2/d}.$$

As $\mathrm{TV}(P_0, P_1) \leq 1$, this can only be done when $f_{\min}\delta^d \lesssim 1$, i.e. $n \gtrsim 1/(f_{\min}\tau^d_{\min})$.

See Appendix G.2 for the construction of these hypotheses. We are now in position to prove Theorem 3.16 (Boundaryless).

Proof of Theorem 3.16 (Boundaryless). Let $\mathcal{P}$ denote the model $\mathcal{P}^{d,D}_{\tau_{\min},\infty}(f_{\min}, f_{\max})$, and write $n_0 := \lceil C_d/(f_{\min}\tau^d_{\min})\rceil$, where $C_d > 0$ is the constant of Proposition F.4.

- If $n \geq n_0$, applying Lemma F.2 (i) with hypotheses P_0 and P_1 of Proposition F.4, yields

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}\left[\mathrm{d_H}\big(M, \hat{M}\big)\right] \geq \frac{1}{2} C'_d \tau_{\min} \left(\frac{1}{f_{\min}\tau_{\min}^d n}\right)^{2/d} \left(1 - \frac{1}{n}\right)^n \geq C''_d \tau_{\min} \left\{1 \wedge \left(\frac{1}{f_{\min}\tau_{\min}^d n}\right)^{2/d}\right\}.$$

- Otherwise, if $n < n_0$, note that since $\inf_{\hat{M}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}\left[\mathrm{d_H}\big(M, \hat{M}\big)\right]$ is a non-increasing sequence, the previous point yields

$$\begin{aligned}\inf_{\hat{M}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}\left[\mathrm{d_H}\big(M, \hat{M}\big)\right] &\geq \inf_{\hat{M}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{n_0}}\left[\mathrm{d_H}\big(M, \hat{M}\big)\right] \\ &\geq C''_d \tau_{\min} \left(\frac{1}{f_{\min}\tau_{\min}^d n_0}\right)^{2/d} \\ &\geq \tilde{C}'_d \tau_{\min} \geq \tilde{C}'_d \tau_{\min} \left\{1 \wedge \left(\frac{1}{f_{\min}\tau_{\min}^d n}\right)^{2/d}\right\},\end{aligned}$$

  which concludes the proof. □

F.2 Convex hypotheses (with boundary)

Similarly to the previous section, we shall use a stability result under diffeomorphisms in the convex case $\tau_{\min} = \infty$. Unfortunately, Proposition F.3 only provides convexity of $\Phi(M)$ (i.e. $\tau_{\Phi(M)} = \infty$) for diffeomorphisms Φ that are affine maps, which does not allow enough flexibility. Beyond affine maps, the following result allows to quantify how much one may bump a strictly convex full dimensional domain while keeping it convex.

Proposition F.5 (Stability of Strict Convexity)**.** *Let $C \subset \mathbb{R}^d$ be a compact domain with $\mathring{C} \neq \emptyset$, that has a $\mathcal{C}^2$ boundary $\bar{\partial}C$. Assume that:*

- *for all $x \in \bar{\partial}C$, $\bar{\partial}C \setminus \{x\}$ is connected;*
- *for all $x, y \in \bar{\partial}C$, $\mathrm{d}(y - x, T_x\bar{\partial}C) \geq A\|y - x\|^2$, for some $A > 0$.*

Let $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ be a $\mathcal{C}^2$ map such that $\lim_{\|x\|\to\infty} \|\Phi(x)\| = \infty$, $\|I_d - d\Phi\|_{\mathrm{op}} \leq 1/10$ and $\left\|d^2\Phi\right\|_{\mathrm{op}} \leq A$, then C and $\Phi(C)$ are convex.

See Appendix G.1 for the proof. Equipped with Propositions F.3 and F.5, we build hypotheses as shown in Figure 11. The formal statement goes as follows.

Proposition F.6 (Convex Hypotheses)**.** *Assume that $f_{\min} \leq c_d/\tau_{\partial,\min}^d$ and $c'_d/\tau_{\partial,\min}^d \leq f_{\max}$ for some small enough $c_d, (c'_d)^{-1} > 0$.*

Then for all $n \geq C_d/(f_{\min}\tau_{\partial,\min}^d)$, there exist $P_0, P_1 \in \mathcal{P}^{d,D}_{\infty,\tau_{\partial,\min}}(f_{\min}, f_{\max})$ with convex supports M_0 and M_1 such that

$$\mathrm{TV}(P_0, P_1) \leq \frac{1}{n} \quad \text{and} \quad \mathrm{d_H}(\partial M_0, \partial M_1) = \mathrm{d_H}(M_0, M_1) \geq C'_d \tau_{\partial,\min} \left(\frac{1}{f_{\min}\tau_{\partial,\min}^d n}\right)^{2/(d+1)}.$$

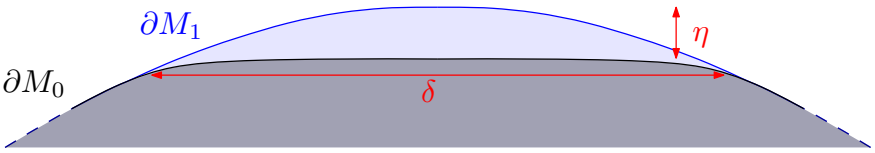


Figure 11: Convex supports M_0 and M_1 of Proposition F.6 for $d = D = 2$. Here, the total variation between the associated uniform distributions is of order $\mathrm{TV}(P_0, P_1) \asymp f_{\min}\mathcal{H}^d(M_0 \triangle M_1) \asymp f_{\min}\delta^{d-1}\eta$ and Hausdorff distance $\mathrm{d_H}(M_0, M_1) = \mathrm{d_H}(\partial M_0, \partial M_1) = \eta$. The reach bound forces the bump to have height $\eta \lesssim \delta^2/\tau_{\partial,\min}$, so that optimal parameter choices yield:

$$\delta \asymp \left(\frac{1}{\tau_{\partial,\min} f_{\min} n}\right)^{1/(d+1)} \quad \text{and} \quad \eta \asymp \frac{\delta^2}{\tau_{\partial,\min}} \asymp \tau_{\partial,\min}\left(\frac{1}{f_{\min}\tau_{\partial,\min}^d n}\right)^{2/(d+1)}.$$

As $\mathrm{TV}(P_0, P_1) \leq 1$, this can only be done when $f_{\min}\delta^{d-1}\eta \lesssim 1$, i.e. $n \gtrsim 1/(f_{\min}\tau_{\partial,\min}^d)$.

See Appendix G.2 for the construction of these hypotheses. We are finally in position to prove Theorem 3.13 and Theorem 3.16 (Convex).

Proofs of Theorem 3.13 and Theorem 3.16 (Convex). The proof follows the lines of that of Theorem 3.16 (Boundaryless) mutatis mutandis. That is, by setting $\mathcal{P} := \mathcal{P}^{d,D}_{\infty,\tau_{\partial,\min}}(f_{\min}, f_{\max})$, $n_0 := \lceil C_d/(f_{\min}\tau_{\partial,\min}^d)\rceil$ where $C_d > 0$ is the constant of Proposition F.6, and applying Lemma F.2 (i) and (ii) with the hypotheses P_0 and P_1 of Proposition F.6. □

G Main tools for the minimax lower bounds

G.1 Stability of the model

G.1.1 Reach bounds

To prove Proposition F.3, we will use the following general reach stability result.

Lemma G.1 ([27, Theorem 4.19])**.** *Let $S \subset \mathbb{R}^D$ with $\tau_S \geq \tau_0 > 0$, and $\Phi : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ be a $\mathcal{C}^1$-diffeomorphism such that Φ,Φ^{-1}, and $d\Phi$ are Lipschitz, with Lipschitz constants K,N and R respectively, then*

$$\tau_{\Phi(S)} \geq \frac{\tau_0}{(K + R\tau_0)N^2}.$$

Proposition F.3 (Reach Stability)**.** *Let $M \in \mathcal{M}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}$ and $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ be a $\mathcal{C}^2$ map such that $\lim_{\|x\|\to\infty} \|\Phi(x)\| = \infty$. Assume that $\sup_{x\in\mathbb{R}^D} \|I_D - d_x\Phi\|_{\mathrm{op}} \leq 1/10$. Then Φ is a global diffeomorphism, and the image $\Phi(M)$ of M by Φ satisfies:*

- $\partial\Phi(M) = \Phi(\partial M)$,
- *If $\sup_{x\in\mathbb{R}^D} \left\|d_x^2\Phi\right\|_{\mathrm{op}} \leq 1/(2\tau_{\min})$, then $\tau_{\Phi(M)} \geq \tau_{\min}/2$,*
- *If $\sup_{x\in\mathbb{R}^D} \left\|d_x^2\Phi\right\|_{\mathrm{op}} \leq 1/(2\tau_{\partial\min})$, then $\tau_{\partial\Phi(M)} \geq \tau_{\partial,\min}/2$.*

Proof of Proposition F.3. First note that since $\sup_x \|d_x\Phi - I_D\|_{\mathrm{op}} < 1$, $d_x\Phi$ is invertible for all $x \in \mathbb{R}^D$, so that Φ is a local diffeomorphism in the neighborhood of x. In addition, $\lim_{\|x\|\to\infty} \|\Phi(x)\| = \infty$, so that the Hadamard-Cacciopoli theorem [21] asserts that Φ is a global diffeomorphism of $\mathbb{R}^D$.

Now, for short, let us write $M' = \Phi(M)$. As Φ is a global diffeomorphism of $\mathbb{R}^D$, M' is a d-dimensional submanifold: indeed, using notation of Definition 2.1, any local $\mathcal{C}^2$ parametrization Ψ_p of M at $p \in M$ lifts to the local $\mathcal{C}^2$ parametrization $\tilde{\Psi}_{\Phi(p)} = \Phi \circ \Psi_p$ of M' at $\Phi(p) \in M'$. In particular, $\partial M' = \Phi(\partial M)$. Moreover, Φ is $\|d\Phi\|_{op} \leq (1 + \|I_D - d\Phi\|_{op})$-Lipschitz, Φ^{-1} is $\left\|d\Phi^{-1}\right\|_{op} \leq (1 - \|I_D - d\Phi\|_{op})^{-1}$-Lipschitz, and $d\Phi$ is $\left\|d^2\Phi\right\|_{op}$-Lipschitz. Hence, Lemma G.1 applied with $S = M$ yields

$$\tau_{M'} \geq \frac{\tau_M (1 - \|I_D - d\Phi\|_{\mathrm{op}})^2}{\|d^2\Phi\|_{op}\, \tau_M + (1 + \|I_D - d\Phi\|_{op})} \geq \tau_M/2 \geq \tau_{\min}/2,$$

where the second inequality used that $\|I_D - d\Phi\|_{\mathrm{op}} \leq 1/10$ and $\left\|d^2\Phi\right\|_{\mathrm{op}} \tau_M \leq 1/2$. Similarly, if the boundary $S = \partial M$ is not empty and $\left\|d^2\Phi\right\|_{\mathrm{op}} \tau_{\partial M} \leq 1/2$, we get

$$\tau_{\partial M'} = \tau_{\Phi(\partial M)} \geq \tau_{\partial M}/2 \geq \tau_{\partial,\min}/2,$$

and otherwise $\tau_{\partial M'} = \tau_\emptyset = \infty \geq \tau_{\partial,\min}/2$, which concludes the proof. □

G.1.2 Strict convexity

To prove Proposition F.5, we will use the following non-standard characterization of convexity for full-dimensional domains.

Lemma G.2. *Let $C \subset \mathbb{R}^d$ be a compact domain with $\mathring{C} \neq \emptyset$, that has a $\mathcal{C}^2$ boundary $\bar{\partial}C$. Assume that:*

- *for all $x \in \bar{\partial}C$, $\bar{\partial}C \setminus \{x\}$ is connected;*
- *for all $x, y \in \bar{\partial}C$, $\mathrm{d}(y - x, T_x\bar{\partial}C) > 0$ as soon as $x \neq y$.*

Then C is convex.

Proof of Lemma G.2. Let us prove the contrapositive. To this aim, assume that C is not convex, meaning that $\tau_C < \infty$. We will prove the existence of points $x, \tilde{y} \in \bar{\partial}C$ such that $\mathrm{d}(\tilde{y} - x, T_x\bar{\partial}C) = 0$.

From [27, Theorem 4.18], there exist $x \neq y \in C$ such that $\mathrm{d}(y - x, Tan(x, C)) > 0$. But for all $x \in \mathring{C}$, $Tan(x, C) = \mathbb{R}^d$, so that $x \in \bar{\partial}C$ necessarily. From here, Proposition 2.6 asserts that $Tan(x, C)$ is a half-space with $\mathrm{span}(Tan(x, C)) = \mathbb{R}^d = T_x\bar{\partial}C \overset{\perp}{\oplus} \mathrm{span}(\eta_x)$ and $Tan(x, C) = \{\langle \eta_x, . \rangle \leq 0\}$, for some unit vector $\eta_x \in \mathbb{R}^d$. Using this representation, for all $z \in C$, we have $\mathrm{d}(z - x, Tan(x, C)) = \langle z - x, \eta_x \rangle_+$ and $\mathrm{d}(z - x, T_x\bar{\partial}C) = |\langle z - x, \eta_x \rangle|$.

On one hand, we have seen that the continuous map $C \ni y \mapsto \langle y - x, \eta_x \rangle_+$ takes a positive value. Hence, by compactness of C, it attains its maximum at some $y_0 \in C$ with $\langle y_0 - x, \eta_x \rangle_+ = \langle y_0 - x, \eta_x \rangle > 0$. But for $\delta \in \mathbb{R}^d$ small enough, $\langle y_0 + \delta - x, \eta_x \rangle_+ = \langle y_0 + \delta - x, \eta_x \rangle = \langle y_0 - x, \eta_x \rangle + \langle \delta, \eta_x \rangle$, so y_0 must belong to $\bar{\partial}C$ as otherwise, y_0 would belong to $\mathring{C}$ and one could increase the value of $\langle \cdot - x, \eta_x \rangle_+$ locally around y_0 and still stay in C.

On the other hand, if we assumed that for all $y \in \bar{\partial}C$, $\langle y - x, \eta_x \rangle \geq 0$ this would lead to a contradiction. Indeed, this inequality would extend to all the points $z \in C$: since C is compact, for all $z \in \mathring{C}$ and $v \in \mathbb{R}^d \setminus \{0\}$, $\{z + \lambda v, \lambda \in \mathbb{R}\} \cap C$ is a non-empty compact set, so there exist $\lambda_- < \lambda_+$ such that for all $\lambda \in [\lambda_-, \lambda_+]^c$, $z + \lambda v \notin C$ and $y_\pm = z + \lambda_\pm v \in C$. In particular, $y_\pm \in \bar{\partial}C$ and $z \in [y_-, y_+] \subset \mathbb{R}^d$. This shows that $z \in C$ can be written as linear combination of elements $y_\pm \in \bar{\partial}C$ and as a result the assumption $\langle y_\pm - x, \eta_x \rangle \geq 0$ would yield $\langle z - x, \eta_x \rangle \geq 0$. This is a contradiction, since by definition of $Tan(x, C) \ni -\eta_x$ (Definition 2.5), there exists $\tilde{z} \in C \setminus \{x\}$ such

that $\left\| -\eta_x - \frac{\tilde{z}-x}{\|\tilde{z}-x\|} \right\| < \frac{1}{2}$ and in particular, $\langle \tilde{z}-x, \eta_x \rangle < 0$. This ends proving that there exists $y_1 \in \bar{\partial} C$ such that $\langle y_1 - x, \eta_x \rangle < 0$.

Summing everything up, we have shown that the continuous map $\bar{\partial} C \setminus \{x\} \ni y \mapsto \langle y-x, \eta_x \rangle$ takes both a positive and a negative value on its connected domain $\bar{\partial} C \setminus \{x\}$. Hence, it must vanish at some point $\tilde{y} \in \bar{\partial} C \setminus \{x\}$, meaning that $x \neq \tilde{y} \in \bar{\partial} C$ and $\mathrm{d}(y-x, T_x \bar{\partial} C) = 0$, which concludes the proof. □

Proposition F.5 (Stability of Strict Convexity)**.** *Let $C \subset \mathbb{R}^d$ be a compact domain with $\mathring{C} \neq \emptyset$, that has a $\mathcal{C}^2$ boundary $\bar{\partial} C$. Assume that:*

- *for all $x \in \bar{\partial} C$, $\bar{\partial} C \setminus \{x\}$ is connected;*
- *for all $x, y \in \bar{\partial} C$, $\mathrm{d}(y-x, T_x \bar{\partial} C) \geq A \|y-x\|^2$, for some $A > 0$.*

Let $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ be a $\mathcal{C}^2$ map such that $\lim_{\|x\| \to \infty} \|\Phi(x)\| = \infty$, $\|I_d - d\Phi\|_{\mathrm{op}} \leq 1/10$ and $\left\|d^2\Phi\right\|_{\mathrm{op}} \leq A$, then C and $\Phi(C)$ are convex.

Proof of Proposition F.5. First, from Lemma G.2, we get that C is convex. Furthermore, as in the proof of Proposition F.3, note that the assumptions $\|d\Phi - I_d\|_{\mathrm{op}} < 1$ and $\lim_{\|x\| \to \infty} \|\Phi(x)\| = \infty$ yield that Φ is a global diffeomorphism of $\mathbb{R}^d$, using the Hadamard-Cacciopoli theorem [21]. Hence, writing $C' = \Phi(C)$, we get that C' is a compact domain with $\mathring{C'} \neq \emptyset$, that has a connected $\mathcal{C}^2$ boundary $\bar{\partial} C'$. In addition, $\bar{\partial} C' = \Phi(\bar{\partial} C)$ and for all $x' = \Phi(x) \in \bar{\partial} C'$, $T_{x'} \bar{\partial} C' = d_x \Phi\big(T_x \bar{\partial} C\big)$.

Now, for all $x, y \in \bar{\partial} C$ and $u \in T_x \bar{\partial} C$, Taylor's theorem and the assumption $\mathrm{d}(y-x, T_x \bar{\partial} C) \geq A \|y-x\|^2$ yield

$$\begin{aligned}
\|d_x\Phi.(y-x) - d_x\Phi.u\| &\geq \left\|d\Phi^{-1}\right\|_{\mathrm{op}}^{-1} \|(y-x)-u\| \\
&\geq \left\|d\Phi^{-1}\right\|_{\mathrm{op}}^{-1} \mathrm{d}(y-x, T_x\bar{\partial}C) \\
&\geq \left\|d\Phi^{-1}\right\|_{\mathrm{op}}^{-1} A \|y-x\|^2 \\
&\geq (1 - \|I_d - d\Phi\|_{\mathrm{op}}) A \|y-x\|^2 \\
&\geq (9A/10) \|y-x\|^2 .
\end{aligned}$$

At second order, Taylor's theorem writes

$$\|\Phi(y) - \Phi(x) - d_x\Phi.(y-x)\| \leq \left\|d^2\Phi\right\|_{\mathrm{op}} \|y-x\|^2 /2.$$

As a result, for all $x' \neq y' \in \bar{\partial} C'$, writing $x' = \Phi(x)$ and $y' = \Phi(y)$ we have $x \neq y$ as Φ^{-1} is one-to-one, and

$$\begin{aligned}
\mathrm{d}(y'-x', T_{x'}\bar{\partial}C') &= \inf_{u \in T_x \bar{\partial} C} \|\Phi(y) - \Phi(x) - d_x\Phi.u\| \\
&\geq \inf_{u \in T_x \bar{\partial} C} \{\|d_x\Phi.(y-x) - d_x\Phi.u\| - \|\Phi(y) - \Phi(x) - d_x\Phi.(y-x)\|\} \\
&\geq \left(9A/10 - \left\|d^2\Phi\right\|_{\mathrm{op}} /2\right) \|y-x\|^2 \\
&> 0,
\end{aligned}$$

since $\left\|d^2\Phi\right\|_{\mathrm{op}} \leq A < 9A/5$. From Lemma G.2, C' is hence convex. □

G.2 Construction of hypotheses

Throughout this section, we will use a smooth localizing bump-type function $\phi : \mathbb{R}^D \to \mathbb{R}$ to build local variations of manifolds. The following result gathers differential estimates, and can be shown using elementary differential calculus.

Proposition G.3. *The localizing function defined as*

$$\begin{aligned}\phi : \mathbb{R}^D &\longrightarrow \mathbb{R} \\ x &\longmapsto \exp\left(-\|x\|^2/(1-\|x\|^2)\right) \mathbb{1}_{\mathrm{B}(0,1)}(x)\end{aligned}$$

is $\mathcal{C}^\infty$ smooth, equal to 0 outside $\mathrm{B}(0,1)$, satisfies $0 \leq \phi \leq 1$, $\phi(0) = 1$,

$$\|d\phi\|_{\mathrm{op}} := \sup_{x\in\mathbb{R}^D} \|d_x\phi\|_{\mathrm{op}} \leq 5/2 \text{ and } \left\|d^2\phi\right\|_{\mathrm{op}} := \sup_{x\in\mathbb{R}^D} \left\|d_x^2\phi\right\|_{\mathrm{op}} \leq 23.$$

G.2.1 Hypotheses with empty boundary

The proof of Proposition F.4 follows that of [3, Lemma 5], and provides a result similar to [31, Theorem 6] in essence. We include it below for sake of completeness and to keep track of explicit constants.

Proposition F.4 (Hypotheses with Empty Boundary)**.** *Assume that $f_{\min} \leq c_d/\tau_{\min}^d$ and $c'_d/\tau_{\min}^d \leq f_{\max}$, for some small enough $c_d, (c'_d)^{-1} > 0$.*

If $d \leq D-1$, then for all $n \geq C_d/(f_{\min}\tau_{\min}^d)$, there exist $P_0, P_1 \in \mathcal{P}^{d,D}_{\tau_{\min},\infty}(f_{\min}, f_{\max})$ with boundariless supports M_0 and M_1 such that

$$\mathrm{TV}(P_0, P_1) \leq \frac{1}{n} \quad \text{and} \quad \mathrm{d_H}(M_0, M_1) \geq C'_d\tau_{\min}\left(\frac{1}{f_{\min}\tau_{\min}^d n}\right)^{2/d}.$$

Proof of Proposition F.4. We let $R = 2\tau_{\min}$, and $M_0 = \mathcal{S}^d(0,R) \times \{0\}^{D-d-1}$ be a d-dimensional sphere of radius R embedded in $\mathbb{R}^{d+1} \times \{0\}^{D-(d+1)}$. Clearly, $\partial M_0 = \emptyset$ (meaning that $\tau_{\partial M_0} = \infty$) and $\tau_{M_0} = R = 2\tau_{\min}$.

Let $e_1 = (1, 0, \ldots, 0)$ denote the first vector of the canonical basis of $\mathbb{R}^D$, and $x_0 = Re_1 \in M_0$. For $\delta > 0$ to be specified later, consider the probability distribution P_0 having the following density with respect to the d-dimensional Hausdorff measure $\mathcal{H}^d$:

$$f_0(x) = 2f_{\min}\mathbb{1}_{M_0\cap\mathrm{B}(x_0,\delta)}(x) + \frac{1-2f_{\min}\mathcal{H}^d(M_0\cap\mathrm{B}(x_0,\delta))}{\mathcal{H}^d(M_0\cap\mathrm{B}(x_0,\delta)^c)}\mathbb{1}_{M_0\cap\mathrm{B}(x_0,\delta)^c}(x),$$

for all $x \in \mathbb{R}^D$. Clearly, P_0 has support M_0 as soon as $2f_{\min}\mathcal{H}^d(M_0 \cap \mathrm{B}(x_0,\delta)) < 1$. In addition, writing σ_d for the volume of the d-dimensional unit Euclidean sphere,

$$\begin{aligned}\frac{1-2f_{\min}\mathcal{H}^d(M_0\cap\mathrm{B}(x_0,\delta))}{\mathcal{H}^d(M_0\cap\mathrm{B}(x_0,\delta)^c)} &\geq \frac{1-2f_{\min}\mathcal{H}^d(M_0\cap\mathrm{B}(x_0,\delta))}{\mathcal{H}^d(M_0)} \\ &= \frac{1-2f_{\min}R^d\mathcal{H}^d\left(\mathrm{B}_{\mathcal{S}^d}(0, 2\arcsin(\delta/(2R)))\right)}{\sigma_d R^d} \\ &\geq \frac{1}{\sigma_d R^d} - 2f_{\min}\left(\frac{\delta}{R}\right)^d.\end{aligned}$$

As a result, $f_0 \geq 2f_{\min}$ over M_0 as soon as $(\sigma_d(2\tau_{\min})^d)^{-1} \geq 4f_{\min}$ and $\delta \leq 2\tau_{\min}$. To upper bound f_0 on M_0, we note that $2f_{\min} \leq f_{\max}/2$ as soon as $2c_d \leq c'_d/2$, and that similarly to above, we derive

$$\frac{1 - 2f_{\min}\mathcal{H}^d(M_0 \cap \mathrm{B}(x_0, \delta))}{\mathcal{H}^d(M_0 \cap \mathrm{B}(x_0, \delta)^c)} \leq \frac{1}{\sigma_d(R^d - \delta^d)} \leq \frac{2}{\sigma_d R^d}$$

as soon as $\delta \leq \tau_{\min}$, which is further upper bounded by $f_{\max}/2$ as soon as $2/(2^d\sigma_d) \leq c'_d/2$. This ends proving that $P_0 \in \mathcal{P}^{d,D}_{\tau_{\min},\infty}(2f_{\min}, f_{\max}/2)$.

We now build P_1 by small and smooth ambient perturbation of P_0. Namely, for $\eta > 0$ to be specified later, write

$$\Phi(x) = x + \eta\phi\left(\frac{x - x_0}{\delta}\right) e_1,$$

where $\phi : \mathbb{R}^D \to \mathbb{R}$ is the localizing function of Proposition G.3. We let $P_1 = \Phi_* P_0$ be the pushforward distribution of P_0 by Φ, and $M_1 = \mathrm{Supp}(P_1)$.

From Proposition G.3, we get that Φ is $\mathcal{C}^\infty$ smooth, $\|d\Phi - I_D\|_{\mathrm{op}} = \frac{\eta}{\delta}\|d\phi\|_{\mathrm{op}} \leq \frac{5\eta}{2\delta}$, and $\left\|d^2\Phi\right\|_{\mathrm{op}} = \frac{\eta}{\delta^2}\left\|d^2\phi\right\|_{\mathrm{op}} \leq \frac{23\eta}{\delta^2}$. Recalling that $\tau_{M_0} \geq 2\tau_{\min}$, Proposition F.3 asserts that $M_1 \in \mathcal{M}^{d,D}_{\tau_{\min},\infty}$ as soon as $\frac{5\eta}{2\delta} \leq \frac{1}{10}$ and $\frac{23\eta}{\delta^2} \leq \frac{1}{4\tau_{\min}}$. Furthermore, from [3, Appendix, Lemma A.6], P_1 admits a density f_1 with respect to $\mathcal{H}^d$ that satisfies

$$f_{\min} = \inf_{M_0} f_0/2 \leq \inf_{M_1} f_1 \leq \sup_{M_1} f_1 \leq 2\sup_{M_0} f_0 \leq f_{\max}$$

as soon as $\frac{5\eta}{2\delta} \leq \frac{1}{3d} \wedge \frac{1}{3(2^{d/2}-1)}$. Hence, under all the above requirements, we finally get that $P_1 \in \mathcal{P}^{d,D}_{\tau_{\min},\infty}(f_{\min}, f_{\max})$.

Now, notice that by construction, $x_0 + \eta e_1 = \Phi(x_0)$ belongs to $M_1 = \Phi(M_0)$. As a result,

$$\mathrm{d_H}(M_0, M_1) \geq \mathrm{d}(x_0 + \eta e_1, M_0) = \eta.$$

In addition, under the same requirements on δ and η as above, Φ is a global diffeomorphism of $\mathbb{R}^D$ (Proposition F.3). As it coincides with the identity map on $\mathrm{B}(x_0, \delta)^c$, this implies that P_0 and $P_1 = \Phi_* P_0$ coincide outside $\mathrm{B}(x_0, \delta)$. Hence,

$$\begin{aligned}
\mathrm{TV}(P_0, P_1) &= \sup_{A \in \mathcal{B}(\mathbb{R}^D)} |P_1(A \cap \mathrm{B}(x_0, \delta)) - P_0(A \cap \mathrm{B}(x_0, \delta))| \\
&\leq \sup_{A \in \mathcal{B}(\mathbb{R}^D)} P_0(A \cap \mathrm{B}(x_0, \delta)) \vee P_1(A \cap \mathrm{B}(x_0, \delta)) \\
&\leq P_0(\mathrm{B}(x_0, \delta)) \vee P_1(\mathrm{B}(x_0, \delta)) \\
&= P_0(\mathrm{B}(x_0, \delta)) \\
&= 2f_{\min}\mathcal{H}^d\big(M_0 \cap \mathrm{B}(x_0, \delta)\big) \\
&= 2f_{\min}R^d\mathcal{H}^d\left(\mathrm{B}_{\mathcal{S}^d}(0, 2\arcsin(\delta/(2R)))\right) \\
&\leq 2\sigma_d f_{\min}\delta^d.
\end{aligned}$$

Setting $2\sigma_d f_{\min}\delta^d = 1/n$ and $\eta = \frac{\delta}{2^{d+10}} \wedge \frac{\delta^2}{92\tau_{\min}}$ (which satisfy all the above requirements) then yields the result, since with that choice, $\delta \leq \tau_{\min}$ and $\eta = \frac{\delta^2}{92\tau_{\min}}$ as soon as $n \geq C_d/(f_{\min}\tau^d_{\min})$ for some large enough $C_d > 0$. □

G.2.2 Convex hypotheses (with boundary)

The proof of Proposition F.6 is similar to that of Proposition F.4.

Proposition F.6 (Convex Hypotheses)**.** *Assume that* $f_{\min} \leq c_d/\tau_{\partial,\min}^d$ *and* $c'_d/\tau_{\partial,\min}^d \leq f_{\max}$ *for some small enough* $c_d, (c'_d)^{-1} > 0$.

Then for all $n \geq C_d/(f_{\min}\tau_{\partial,\min}^d)$*, there exist* $P_0, P_1 \in \mathcal{P}_{\infty,\tau_{\partial,\min}}^{d,D}(f_{\min}, f_{\max})$ *with convex supports* M_0 *and* M_1 *such that*

$$\mathrm{TV}(P_0, P_1) \leq \frac{1}{n} \quad \text{and} \quad \mathrm{d_H}(\partial M_0, \partial M_1) = \mathrm{d_H}(M_0, M_1) \geq C'_d \tau_{\partial,\min} \left(\frac{1}{f_{\min}\tau_{\partial,\min}^d n} \right)^{2/(d+1)}.$$

Proof of Proposition F.6. Let $R = 2\tau_{\partial,\min}$, and $M_0 = \mathrm{B}_{\mathbb{R}^d}(0, R) \times \{0\}^{D-d}$ be a d-dimensional ball of radius R embedded in $\mathbb{R}^d \times \{0\}^{D-d}$. Clearly, M_0 is convex, meaning that $\tau_{M_0} = \infty$, and $\partial M_0 = \mathcal{S}^{d-1}(0, R) \times \{0\}^{D-d}$ has reach $\tau_{\partial M_0} = R$.

Let $e_1 = (1, 0, \ldots, 0)$ denote the first vector of the canonical basis of $\mathbb{R}^D$, and $x_0 = Re_1 \in M_0$. For $\delta > 0$ to be specified later, consider the probability distribution P_0 having the following density with respect to the d-dimensional Hausdorff measure $\mathcal{H}^d$:

$$f_0(x) = 2f_{\min} \mathbb{1}_{M_0 \cap \mathrm{B}(x_0,\delta)}(x) + \frac{1 - 2f_{\min}\mathcal{H}^d(M_0 \cap \mathrm{B}(x_0, \delta))}{\mathcal{H}^d(M_0 \cap \mathrm{B}(x_0, \delta)^c)} \mathbb{1}_{M_0 \cap \mathrm{B}(x_0,\delta)^c}(x),$$

for all $x \in \mathbb{R}^D$. We see that P_0 has support M_0 if $2f_{\min}\mathcal{H}^d(M_0 \cap \mathrm{B}(x_0, \delta)) < 1$. Denoting by ω_d the volume of the d-dimensional unit Euclidean ball, we derive

$$\begin{aligned}
\frac{1 - 2f_{\min}\mathcal{H}^d(M_0 \cap \mathrm{B}(x_0, \delta))}{\mathcal{H}^d(M_0 \cap \mathrm{B}(x_0, \delta)^c)} &\geq \frac{1 - 2f_{\min}\mathcal{H}^d(M_0 \cap \mathrm{B}(x_0, \delta))}{\mathcal{H}^d(M_0)} \\
&\geq \frac{1 - 2f_{\min}(\omega_d \delta^d/2)}{\omega_d R^d} \\
&\geq \frac{1}{\omega_d R^d} - f_{\min} \left(\frac{\delta}{R} \right)^d.
\end{aligned}$$

As a result, $f_0 \geq 2f_{\min}$ over M_0 as soon as $(\omega_d(2\tau_{\min})^d)^{-1} \geq 4f_{\min}$ and $\delta \leq 2\tau_{\min}$, To upper bound f_0 on M_0, we note that $2f_{\min} \leq f_{\max}/2$ as soon as $2c_d \leq c'_d/2$, and that similarly to above, we derive

$$\frac{1 - 2f_{\min}\mathcal{H}^d(M_0 \cap \mathrm{B}(x_0, \delta))}{\mathcal{H}^d(M_0 \cap \mathrm{B}(x_0, \delta)^c)} \leq \frac{1}{\omega_d(R^d - \delta^d/2)} \leq \frac{2}{\omega_d R^d}$$

as soon as $\delta \leq R = 2\tau_{\min}$, which is further upper bounded by $f_{\max}/2$ as soon as $2/(2^d\omega_d) \leq c'_d/2$. In all, we have $P_0 \in \mathcal{P}_{\infty,\tau_{\partial,\min}}^{d,D}(2f_{\min}, f_{\max}/2)$.

Now, to build P_1, let $\eta > 0$ be a parameter to be specified later, and write

$$\Phi(x) = x + \eta\phi\left(\frac{x - x_0}{\delta} \right) e_1,$$

where $\phi : \mathbb{R}^D \to \mathbb{R}$ is the localizing function of Proposition G.3. We let $P_1 = \Phi_* P_0$ be the pushforward distribution of P_0 by Φ, and $M_1 = \mathrm{Supp}(P_1)$. Note by now that if $\delta \leq R$, we have $M_0 \subset M_1$.

From Proposition G.3, we get that Φ is $\mathcal{C}^\infty$ smooth, $\|d\Phi - I_D\|_{\mathrm{op}} = \frac{\eta}{\delta}\|d\phi\|_{\mathrm{op}} \leq \frac{5\eta}{2\delta}$, and $\|d^2\Phi\|_{\mathrm{op}} = \frac{\eta}{\delta^2}\|d^2\phi\|_{\mathrm{op}} \leq \frac{23\eta}{\delta^2}$. It is also clear that $\lim_{\|x\|\to\infty}\|\Phi(x)\| = \infty$. Hence, recalling that $\tau_{\partial M_0} \geq 2\tau_{\partial,\min}$, Proposition F.3 asserts that $\tau_{\partial M_1} \geq \tau_{\partial,\min}$ as soon as $\frac{5\eta}{2\delta} \leq \frac{1}{10}$ and $\frac{23\eta}{\delta^2} \leq \frac{1}{4\tau_{\partial,\min}}$. In addition, as Φ preserves $\mathbb{R}^d \times \{0\}^{D-d}$, both M_0 and M_1 can be seen as compact domains of $\mathbb{R}^d$ with non-empty interior. In this d-plane $\mathbb{R}^d \times \{0\}^{D-d} \cong \mathbb{R}^d$, M_0 has a $\mathcal{C}^2$ (topological) boundary $\bar{\partial}M_0 = \mathcal{S}^{d-1}(0,R)$, the set $\bar{\partial}M_0 \setminus \{x\}$ is connected for all $x \in \bar{\partial}M_0$ (note that for $d=1$, this set is only reduced to a point), and for all $x, y \in \bar{\partial}M_0$, $\mathrm{d}(y-x, T_x\bar{\partial}M_0) = \frac{1}{4\tau_{\partial,\min}}\|y-x\|^2$. As a result, Proposition F.5 applied with $k=d$ asserts that $M_1 = \Phi(M_0)$ remains convex as soon as $\frac{5\eta}{2\delta} \leq \frac{1}{10}$ and $\frac{23\eta}{\delta^2} \leq \frac{1}{4\tau_{\partial,\min}}$. This ends proving that $M_0, M_1 \in \mathcal{M}^{d,D}_{\infty,\tau_{\partial,\min}}$ under the above requirements.

Furthermore, from [3, Appendix, Lemma A.6], we get that P_1 admits a density f_1 with respect to $\mathcal{H}^d$ that satisfies

$$f_{\min} = \inf_{M_0} f_0/2 \leq \inf_{M_1} f_1 \leq \sup_{M_1} f_1 \leq 2\sup_{M_0} f_0 \leq f_{\max}$$

as soon as $\frac{5\eta}{2\delta} \leq \frac{1}{3d} \wedge \frac{1}{3(2^{d/2}-1)}$. Hence, under all the above requirements, we have that both P_0 and P_1 belong to the model $\mathcal{P}^{d,D}_{\infty,\tau_{\partial,\min}}(f_{\min}, f_{\max})$.

Further analyzing the properties of f_1, let $y \in M_1 \cap \mathrm{B}(x_0,\delta)$. As the diffeomorphism Φ maps $\mathrm{B}(x_0,\delta)$ onto itself, $y = \Phi(x)$ for a unique $x \in M_0 \cap \mathrm{B}(x_0,\delta)$. Hence, applying [3, Appendix, Lemma A.6] again we get

$$\begin{aligned}
|f_1(y) - 2f_{\min}| &= |f_1(y) - f_0(x)| \\
&\leq f_0(x)\left(\frac{3d}{2} \vee 3(2^{d/2}-1)\right)\|d\Phi - I_D\|_{\mathrm{op}} \\
&= \frac{2^{d+10} f_{\min}\eta}{\delta},
\end{aligned}$$

provided that $\frac{5\eta}{2\delta} < \frac{1}{3}$. From this bound, we also read that $f_1 \leq 3f_{\min}$ on $M_1 \cap \mathrm{B}(x_0,\delta)$ as soon as $\frac{2^{d+10}\eta}{\delta} \leq 1$. We can now move forward and prove the result.

First, notice that by construction $x_0 + \eta e_1 = \Phi(x_0)$ belongs to $\partial M_1 = \Phi(\partial M_0)$. As a result,

$$\mathrm{d_H}(\partial M_0, \partial M_1) = \mathrm{d_H}(M_0, M_1) \geq \mathrm{d}(x_0 + \eta e_1, M_0) = \eta.$$

Second, under the same requirements on δ and η as above, Φ is a global diffeomorphism of $\mathbb{R}^D$ (Proposition F.3). As it coincides with the identity map on $\mathrm{B}(x_0,\delta)^c$, it implies that P_0 and $P_1 = \Phi_* P_0$ coincide outside $\mathrm{B}(x_0,\delta)$. Applying the second formula of Definition F.1 with the σ-finite dominating measure $\mu = \mathbb{1}_{\mathbb{R}^d\times\{0\}^{D-d}}\mathcal{H}^d$, we hence get

$$\begin{aligned}
\mathrm{TV}(P_0, P_1) &= \frac{1}{2}\int_{\mathrm{B}(x_0,\delta)\cap(M_0\cup M_1)} |f_1 - f_0| d\mathcal{H}^d \\
&= \frac{1}{2}\int_{\mathrm{B}(x_0,\delta)\cap M_0} |f_1 - 2f_{\min}| d\mathcal{H}^d + \frac{1}{2}\int_{\mathrm{B}(x_0,\delta)\cap(M_1\setminus M_0)} f_1 d\mathcal{H}^d \\
&\leq \frac{2^{d+10} f_{\min}\eta}{2\delta}\mathcal{H}^d\big(\mathrm{B}(x_0,\delta)\cap M_0\big) + \frac{3f_{\min}}{2}\mathcal{H}^d\big(\mathrm{B}(x_0,\delta)\cap(M_1\setminus M_0)\big).
\end{aligned}$$

Furthermore, by construction, $\mathcal{H}^d\big(\mathrm{B}(x_0,\delta)\cap M_0\big) \leq \omega_d\delta^d/2$ and $\mathcal{H}^d\big(\mathrm{B}(x_0,\delta)\cap(M_1\setminus M_0)\big) \leq C'_d\delta^{d-1}\eta$, so that

$$\mathrm{TV}(P_0, P_1) \leq C''_d f_{\min}\delta^{d-1}\eta.$$

Finally, setting $C''_d f_{\min} \delta^{d-1} \eta = 1/n$ and $\eta = \frac{\delta}{2^{d+10}} \wedge \frac{\delta^2}{92\tau_{\partial,\min}}$ (which satisfy all the above requirements) then yields the result, since with that choice, $\delta \leq \tau_{\min}$ and $\eta = \frac{\delta^2}{92\tau_{\partial,\min}}$ as soon as $n \geq \tilde{C}_d/(f_{\min}\tau^d_{\partial,\min})$ for some large enough $C_d > 0$. □

References

[1] Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the reach of a manifold. *Electron. J. Stat.*, 13(1):1359–1399, 2019.

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

[3] Eddie Aamari and Clément Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *Ann. Statist.*, 47(1):177–204, 2019.

[4] Catherine Aaron and Olivier Bodart. Local convex hull support and boundary estimation. *J. Multivariate Anal.*, 147:82–101, 2016.

[5] Catherine Aaron and Alejandro Cholaquidis. On boundary detection. *Ann. Inst. Henri Poincaré Probab. Stat.*, 56(3):2028–2050, 2020.

[6] Catherine Aaron, Alejandro Cholaquidis, and Ricardo Fraiman. Estimation of surface area. *Electronic Journal of Statistics*, 16(2):3751 – 3788, 2022.

[7] Yariv Aizenbud and Barak Sober. Non-Parametric Estimation of Manifolds from Noisy Data. *arXiv e-prints*, page arXiv:2105.04754, May 2021.

[8] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434, 2006.

[9] Clément Berenfeld, John Harvey, Marc Hoffmann, and Krishnan Shankar. Estimating the reach of a manifold via its convexity defect function. *Discrete & Computational Geometry*, Jun 2021.

[10] Tyrus Berry and Timothy Sauer. Density estimation on manifolds with boundary. *Comput. Statist. Data Anal.*, 107:1–17, 2017.

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

[12] 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.

[13] Jean-Daniel Boissonnat, André Lieutier, and Mathijs Wintraecken. The reach, metric distortion, geodesic convexity and the variation of tangent spaces. *J. Appl. Comput. Topol.*, 3(1-2):29–58, 2019.

[14] Jean-Daniel Boissonnat and Mathijs Wintraecken. The Topological Correctness of PL-Approximations of Isomanifolds. In Sergio Cabello and Danny Z. Chen, editors, *36th International Symposium on Computational Geometry (SoCG 2020)*, volume 164 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 20:1–20:18, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

[15] Glen E. Bredon. *Topology and geometry*, volume 139 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1993.

[16] Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2001.

[17] Jeff Calder, Sangmin Park, and Dejan Slepčev. Boundary estimation from point clouds: Algorithms, guarantees and applications. *Journal of Scientific Computing*, 92(2):56, Jul 2022.

[18] Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Convergence rates for persistence diagram estimation in topological data analysis. *J. Mach. Learn. Res.*, 16:3603–3635, 2015.

[19] Frédéric Chazal and Bertrand Michel. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. *arXiv e-prints*, page arXiv:1710.04019, October 2017.

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

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

[22] Tamal K Dey, Kuiyu Li, Edgar A Ramos, and Rephael Wenger. Isotopic reconstruction of surfaces with boundaries. In *Computer Graphics Forum*, volume 28, pages 1371–1382. Wiley Online Library, 2009.

[23] Vincent Divol. Minimax adaptive estimation in manifold inference. *arXiv e-prints*, page arXiv:2001.04896, January 2020.

[24] Vincent Divol. Reconstructing measures on manifolds: an optimal transport approach. *arXiv e-prints*, page arXiv:2102.07595, February 2021.

[25] Manfredo Perdigão do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser Boston, Inc., Boston, MA, 1992. Translated from the second Portuguese edition by Francis Flaherty.

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

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

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

[29] Charles Fefferman, Sergei Ivanov, Matti Lassas, and Hariharan Narayanan. Fitting a manifold of large reach to noisy data. *arXiv e-prints*, page arXiv:1910.05084, October 2019.

[30] 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.

[31] Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Minimax manifold estimation. *J. Mach. Learn. Res.*, 13:1263–1291, 2012.

[32] Sariel Har-Peled. *Geometric approximation algorithms*, volume 173 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2011.

[33] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

[34] Allen Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge, 2002.

[35] Morris W. Hirsch. *Differential topology*. Graduate Texts in Mathematics, No. 33. Springer-Verlag, New York-Heidelberg, 1976.

[36] 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.

[37] John A. Lee and Michel Verleysen. *Nonlinear dimensionality reduction*. Information Science and Statistics. Springer, New York, 2007.

[38] John M. Lee. *Introduction to topological manifolds*, volume 202 of *Graduate Texts in Mathematics*. Springer, New York, second edition, 2011.

[39] Mauro Maggioni, Stanislav Minsker, and Nate Strawn. Multiscale dictionary learning: non-asymptotic bounds and robustness. *J. Mach. Learn. Res.*, 17:Paper No. 2, 51, 2016.

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

[41] J. Mø ller. Random tessellations in $\mathbf{R}^d$. *Adv. in Appl. Probab.*, 21(1):37–73, 1989.

[42] 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.

[43] Nikita Puchkin and Vladimir Spokoiny. Structure-adaptive manifold estimation. *arXiv e-prints*, page arXiv:1906.05014, June 2019.

[44] Laurent Rineau and Mariette Yvinec. Meshing 3d domains bounded by piecewise smooth surfaces*. In Michael L. Brewer and David Marcum, editors, *Proceedings of the 16th International Meshing Roundtable*, pages 443–460, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[45] Alberto Rodríguez Casal. Set estimation under convexity type assumptions. *Ann. Inst. H. Poincaré Probab. Statist.*, 43(6):763–774, 2007.

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

[47] Donald R Sheehy. An output-sensitive algorithm for computing weighted α-complexes. In *CCCG*, 2015.

[48] Larry Wasserman. Topological data analysis. *Annu. Rev. Stat. Appl.*, 5:501–535, 2018.

[49] Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, New York, 1997.