Convergence Rates for Geometric Inference

Eddie Aamari

NNT : 2017SACLS203

THÈSE DE DOCTORAT

de

L'Université Paris-Saclay

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : Université Paris-Sud

Laboratoire d'accueil : Laboratoire de mathématiques d'Orsay, UMR 8628 CNRS

Spécialité de doctorat : Mathématiques aux interfaces

Eddie AAMARI

Vitesses de convergence en inférence géométrique

Date de soutenance : 1er septembre 2017

Après avis des rapporteurs :
Bruno PELLETIER (Université Rennes II)
Wolfgang POLONIK (University of California, Davis)

Jury de soutenance :

Jean-Daniel BOISSONNAT	(Inria Sophia-Antipolis) Examinateur
Stéphane BOUCHERON	(Université Paris-Diderot) Président du jury
Frédéric CHAZAL	(Inria Saclay) Directeur de thèse
Pascal MASSART	(Université Paris-Saclay) Directeur de thèse
Bruno PELLETIER	(Université Rennes II) Rapporteur
Wolfgang POLONIK	(University of California, Davis) Rapporteur

Remerciements

Avant tout, j'aimerais remercier chaleureusement Frédéric et Pascal pour leur confiance sans faille. Confiance, en premier lieu, pour avoir accepté d'encadrer un étudiant que vous connaissiez peu. Confiance, ensuite, pour ne jamais m'avoir ôté la latitude de prendre autant de temps que nécessaire sur un problème, malgré mes tâtonnements nombreux. Nos échanges m'ont énormément appris, et ce, dans une ambiance toujours studieuse mais légère. C'était un immense plaisir de travailler avec vous.

I would also like to thank warmly and respectfully Bruno Pelletier and Wolfgang Polonik, who accepted to dedicate time to report this thesis.

Jean-Daniel Boissonnat et Stéphane Boucheron ont gentiment accepté de participer à mon jury, malgré leurs éloignements respectifs des côtés statistiques et géométriques de cette thèse. Je tiens à les en remercier.

Cette thèse est essentiellement le fruit de deux collaborations. Il va sans dire que le travail présenté ici n'aurait pas été si riche sans toutes les bonnes idées de ceux avec qui j'ai pu m'associer. Clément, j'ai trouvé nos pérégrinations sans fin très chouettes. On remet ça quand tu veux. Jisu, I really feel lucky for having been able to get started in research with you. Also, Pittsburgh wouldn't have been as enjoyable as it was, without all those restaurants only you know how to find.

It would be a shame to omit the two other Pittsburgh folks I had the chance to work with. Larry and Alessandro, I am very grateful to you for welcoming me as a peer and in such a friendly way. Working with you was a delight. Un remerciement tout particulier, aussi, à Bertrand. Avec un tel degré de disponibilité et de sympathie, tu as dépassé de loin ce que j'aurais pu attendre d'un troisième directeur de thèse.

Travailler au laboratoire d'Orsay et au sein de l'équipe Geometrica/DataShape a été véritablement épanouissant. Merci à ses membres pour leur vivacité, aux administratives pour leur patience, ainsi que pour toutes les discussions passionnantes que l'on a pu avoir.

J'ai aussi eu l'occasion de participer à de nombreux séminaires et colloques durant ces trois années, avec à la clé des rencontres qui m'ont considérablement fait avancer. J'en sortais la plupart du temps avec de belles idées neuves. Un grand merci.

Une pensée, enfin, pour tou-te-s les Charentais-es, Poitevin-e-s, Rennais-es, Francilien-ne-s et autres, qui sont moins attaché-e-s au contenu scientifique de ce manuscrit qu'à son auteur. Vous m'avez construit, cette thèse est donc aussi la vôtre.

Contents

Chapitre I

Introduction générale

La recherche de caractéristiques pertinentes associées à des jeux de données suscite un intérêt croissant du fait de leur acquisition massive. Ces caractéristiques ont pour but de résumer des données non structurées, souvent représentées par des nuages de points en grande dimension, en les réduisant à des descripteurs simples à analyser. Pour appréhender des données vivant en grande dimension, un cadre statistique raisonnable consiste à supposer qu'elles se concentrent sur un ensemble de dimension intrinsèque d, petite par rapport à la dimension D de l'espace des mesures. Ce postulat est fondé sur l'idée que les données contiennent une forme de redondance ou de corrélation, et qu'elles ne comportent pas véritablement D degrés de liberté.

Les techniques linéaires de réduction de dimension ont fait l'objet de nombreux travaux. En particulier, les méthodes parcimonieuses du type LASSO, qui visent à annuler des coefficients d'un paramètre à estimer, ont connu un essor considérable [HTF09]. Au delà des techniques mises en œuvre, notons que l'idée d'annuler des coefficients repose de manière implicite sur la confiance en une paramétrisation particulière du phénomène. En effet, le caractère creux d'un vecteur n'est pas stable par déformation, même rigide, de l'espace ambiant. Lorsqu'un système de coordonnées n'est pas fiable, interprétable, ou pas même disponible — par exemple pour des données prenant la forme d'une matrice de distances [GG12] —, de telles méthodes ne peuvent pas s'appliquer directement. De manière encore plus critique, un modèle de régression peut ne s'appliquer dans aucun système de coordonnées. En effet, les graphes de fonctions modélisent mal des jeux de données repliés sur eux-mêmes, présentant une topologie non triviale autre que celle d'un convexe. C'est par exemple le cas pour les configurations admissibles de cetains systèmes thermodynamiques, les conformations de biomolécules, ou bien la répartition filamentaire des galaxies en cosmologie [LV07]. Les données présentent alors une géométrie dont les caractéristiques peuvent être informatives, donc intéressantes à étudier.

Dans un contexte tout autre, la géométrie algorithmique s'intéresse au traitement des problèmes de nature géométrique sur ordinateur [BY98]. Par exemple, si l'on dispose d'une numérisation discrète d'un objet continu, on peut chercher à le reconstruire avec des triangulations. Dans ce domaine, les guaranties théoriques obtenues reposent sur des conditions déterministes d'échantillonnage, souvent basées sur la densité et la généricité d'un nuage de points vis-à-vis de la forme sous-jacente. En dimension 2 et 3, l'estimation de descripteurs géométriques a déjà été très étudiée. Les nombreuses techniques existantes, asorties d'heuristiques ainsi que de structures de données efficaces fournissent un contexte satisfaisant pour leur utilisation [BT07]. En dimension supérieure, on trouve une littérature assez riche pour l'inférence de caractéristiques topologiques, mais bien moins d'occurrences pour des quantités géométriques.

Géométrie des données

L'analyse topologique de données est un domaine visant à extraire une information de nature géométrique ou topologique à partir de données [Car09]. Les mesures sont considérées comme étant générées sur «une forme M», ce qui ouvre de nombreuses questions sur leur géométrie. On s'intéresse alors à des notions qui sont invariantes par changement de coordonnées. De fait, cela amène aussi à considérer des invariants topologiques et des quantités intrinsèques issues de la géométrie différentielle comme pouvant résumer les données. On peut ensuite tirer profit de ces signatures géométriques en utilisant des méthodes classiques d'apprentissage, par exemple via des techniques de segmentation ou de classification [HTF09]. L'analyse topologique de données se place donc à l'interface entre géométrie algorithmique et statistiques. Elle soulève notamment le problème de la grande dimension en géométrie algorithmique et réciproquement, elle amène de nouveaux descripteurs en statistiques.

Décrivons maintenant quelques objets géométriques et topologiques d'intérêt, ainsi que l'interprétation que l'on peut en avoir vis-à-vis des données étudiées. On adopte le cadre, qui nous intéressera tout au long de cette thèse, dans lequel les données proviennent d'une sous-variété source $M \subset \mathbb{R}^D$ de classe au moins $\mathcal{C}^2$.

D'abord, la sous-variété M elle-même renseigne sur la localisation des données. On parle alors d'estimation de support, ou d'ensemble [Cue09]. La qualité de l'approximation de M par un estimateur $\hat{M}$ peut être évaluée par différentes pertes selon les propriétés de l'estimation recherchée. Parmi les plus classiques, citons la mesure de la différence symétrique $\mu(M \triangle \hat{M})$, qui peut garantir la détection fine de valeurs aberrantes dans la surveillance d'un système [BCP08]. D'autre part, la distance de Hausdorff $d_H(M, \hat{M})$ fournit une mesure assez rigide pour garantir des propriétés de stabilité géométrique dans les cas réguliers [CCSL06]. La dimension intrinsèque $d = \dim(M)$ informe sur le nombre de degrés de liberté du système sous-jacent [LV07]. Inversement, la codimension $D - d$ précise le nombre de corrélations locales entre les variables étudiées. En homologie, le nombre de Betti $\beta_0(M)$ correspond au nombre de composantes connexes de M. Sa détermination préalable peut être nécessaire dans certains algorithmes de clustering [SJ03]. L'espace tangent T_xM est la meilleure approximation linéaire de M en $x \in M$. Par conséquent, il fournit les directions de grande variabilité locale du système étudié [ACLZ17]. C'est aussi un bon candidat pour un domaine de paramétrisation locale de dimension d. Concernant les quantités différentielles d'ordre deux, la seconde forme fondamentale II_x^M précise à quel point M s'éloigne localement du cadre linéaire T_xM, ainsi que les directions dans lesquelles cela a lieu. Elle encode complètement la courbure et informe sur une échelle locale à laquelle regarder les données [CP05]. On peut aussi citer d'autres objets globaux et plus élaborés tels que le reach [DS06], le volume [BH98], les nombres de Betti d'ordres supérieurs $\beta_k(M)$ [BRS+12], la persistance topologique [Oud15], les graphes de Reeb [GSBW11], le bord [CRC04], la distance géodésique [MS05] ou la distance à la mesure [CCSM11]. La liste est bien loin d'être exhaustive [Was]. Chacun de ces objets peut constituer un paramètre d'intérêt attaché aux données que l'on peut chercher à estimer.

Cette thèse s'intéresse à l'estimation optimale, à partir de nuages de points $\mathbb{X}_n = \{X_1, \ldots, X_n\}$, de quantités géométriques associées à des sous-variétés M de l'espace euclidien $\mathbb{R}^D$, dans un cadre statistique non-asymptotique. Nous examinons les vitesses optimales d'estimation de ces objets pour différentes classes de régularité de la sous-variété source inconnue M.

Optimalité et vitesse de convergence

Jusqu'à récemment, les questions d'optimalité ont été peu traitées en inférence géométrique. En géométrie algorithmique, l'optimalité fait souvent référence à la complexité algorithmique lorsque le problème est de nature combinatoire [AL13]. Quand ce n'est pas le cas, la notion d'optimalité repose sur des constructions de nuages de points ad hoc et non génériques [Cla06]. À l'inverse, les notions d'optimalité abondent en statistiques à la fois paramétrique [LC98] et non-paramétrique [Tsy09]. Dans cette thèse, nous utiliserons le risque minimax, un critère d'optimalité très répandu en statistique non-paramétrique. Décrivons maintenant sa construction dans un cadre général.

On considère le problème de l'estimation d'un paramètre d'intérêt $\theta(P)$ dépendant de la loi commune $P \in \mathcal{P}$ d'un n-échantillon $\mathbb{X}_n = \{X_1, \dots, X_n\}$ ayant pour support $M = Supp(P)$. Pour $\theta(P)$, on peut penser par exemple aux quantités géométriques et topologiques décrites auparavant, ou bien de manière plus classique à une fonction de régression ou une densité. On cherche à répondre à la question *«peut-on estimer $\theta(P)$ étant donné un n-échantillon $\mathbb{X}_n = \{X_1, \dots, X_n\}$ de loi P ?»*. Par estimer, on veut dire trouver $\hat{\theta} = \hat{\theta}(\mathbb{X}_n)$ qui rende petite, en moyenne, une mesure de qualité $d(\theta(P), \hat{\theta})$ préalablement fixée. À P fixée, le théorème fondamental de la statistique affirme que la mesure empirique P_n converge presque sûrement vers P quand n tend vers l'infini. Pourvu que la fonctionnelle d'intérêt $P \mapsto \theta(P)$ soit stable par rapport à P, on peut espérer estimer $\theta(P)$ au moins asymptotiquement. Cependant, il est impossible d'obtenir une vitesse de convergence quand n tend vers l'infini si P est autorisée à se rapprocher de cas pathologiques. Par exemple, si aucune hypothèse de régularité n'est faite sur une fonction de régression, il est sans espoir de discerner le signal du bruit.

Pour obtenir une réponse à la question, plus précise, *«à quelle vitesse peut-on estimer $\theta(P)$?»*, il faut donc restreindre le domaine d'étude — un modèle $\mathcal{P}$ réputé contenir P — et en étudier les limitations intrinsèques. Le risque minimax $R_n(\mathcal{P})$ sur le modèle $\mathcal{P}$ pour l'estimation du paramètre $\theta(P)$ sur un échantillon de taille n est le meilleur risque moyen atteignable uniformément sur $\mathcal{P}$ par un estimateur. C'est-à-dire,

$$R_n(\mathcal{P}) = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} d(\theta(P), \hat{\theta}),$$

où l'infimum est pris sur l'ensemble des estimateurs $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$. Le risque minimax correspond à la meilleure performance qu'il est possible d'obtenir à partir de n points. Lorsque n devient grand, $R_n(\mathcal{P})$ informe sur la vitesse optimale d'approximation de la quantité d'intérêt $\theta(P)$ sur $\mathcal{P}$. Par conséquent, un estimateur $\hat{\theta}$ est dit optimal au sens minimax lorsque, pour n assez grand,

$$R_n(\mathcal{P}) \leq \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} d(\theta(P), \hat{\theta}) \leq C_{\mathcal{P}} R_n(\mathcal{P}), \tag{I.1}$$

pour $C_{\mathcal{P}} > 1$. Pour étudier le comportement d'un risque minimax et obtenir un résultat du type (I.1), on procède généralement en deux étapes tout à fait indépendantes.

(i) Borner supérieurement le risque minimax revient à exhiber un estimateur $\hat{\theta}$ et à en étudier la performance uniformément sur $\mathcal{P}$. Cela se résume en une borne $\sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} d(\theta(P), \hat{\theta}) \leq v_n$. Cette borne est spécifique à chaque problème et repose sur les propriétés non-asymptotiques de l'estimateur $\hat{\theta}$ considéré.

(ii) Pour démontrer qu'on ne peut pas faire mieux, une borne inférieure $R_n(\mathcal{P}) \geq v'_n$ est obtenue avec des arguments Bayésiens [Yu97]. L'argument sous-jacent correspond à une étude en pire cas. En effet, si (au moins) deux distributions $P_1, P_2 \in \mathcal{P}$ sont

telles que leurs n-échantillons respectifs ont des lois proches mais que les paramètres $\theta(P_1)$ et $\theta(P_2)$ sont éloignés, aucun estimateur ne peut à la fois être proche de $\theta(P_1)$ et $\theta(P_2)$ simultanément et commettra une erreur de l'ordre de $d(\theta(P_1), \theta(P_2))$ avec grande probabilité.

Si $v_n \leq C_{\mathcal{P}} v'_n$, on obtient alors (I.1) et l'on dit que v_n, ou de manière équivalente v'_n, est la vitesse optimale d'estimation de $\theta(P)$ sur $\mathcal{P}$.

Ici, le fait de considérer un risque moyen sur des n-échantillons permet de formaliser la notion de généricité d'un nuage de points. Ainsi, une procédure renvoyant une quantité $\hat{\theta}(x_1, \ldots, x_n)$ contruite sur n points $\{x_1, \ldots, x_n\}$ (vus comme déterministes) pourra être considérée comme génériquement optimale si elle atteint la vitesse minimax lorsqu'elle est évaluée sur un n-échantillon $\mathbb{X}_n$. L'aléa donne un cadre où la notion d'optimalité est bien posée.

Notions quantitatives de régularité géométrique

Comme décrit précédemment, une étude minimax nécessite la spécification préalable d'un modèle $\mathcal{P}$. Celui-ci définit la classe de régularité des objets étudiés. Ici, la notion de régularité est à prendre au sens large et peut recouvrir la dimension, la massivité d'espace, l'approximabilité des objets, ou le caractère lisse au sens différentiel classique. Plus cette classe est grande, plus le problème est général et difficile. Plus cette classe est restreinte, plus le problème est spécifique et donc accessible. Définir un modèle revient donc à caractériser quantitativement la difficulté d'estimation d'un objet.

Par régularité quantitative, on exprime la nécessité de borner les objets étudiés. En analogie avec la regression, les classes $\mathcal{C}^k$ de fonctions k fois continûment différentiables ne présentent pas une information suffisante pour exploiter de façon uniforme leur régularité. La densité de $\mathcal{C}^k$ dans l'ensemble des fonctions continues atteste du fait qu'il est possible de s'approcher de cas pathologiques non-lisses. À l'inverse, les classes de Hölder $\mathcal{C}^k(L)$ — composées des fonctions $k-1$ fois différentiables dont la dérivée $(k-1)$-ième est L-Lipschitzienne — évitent ce phénomène. De la même manière, pour des modèles de sous-variétés de $\mathbb{R}^D$, la régularité doit être quantifiée. L'absence de système de coordonnées canonique et de paramétrisation naturelle rend le sujet plus complexe que pour les fonctions.

Une première caractéristique de régularité est la dimension intrisèque $d = \dim(M)$. Elle régit en premier lieu la massivité métrique d'une sous-variété M, de la même manière que dans le cas euclidien. Par la suite, on supposera toujours la dimension d connue. Nous verrons que d influe fortement sur les vitesses d'estimation des objets étudiés. Par ailleurs, nous prêterons une attention particulière à développer une analyse insensible à la dimension ambiante D qui, extrinsèque, est potentiellement très grande devant d.

Ensuite, à dimension d fixée, un paramètre de régularité particulièrement populaire en inférence géométrique est le reach[1]. Introduit pour la première fois par Herbert Federer dans le cadre de la théorie géométrique de la mesure [Fed59], le reach τ_M de $M \subset \mathbb{R}^D$ est le plus grand rayon $r \geq 0$ tel que tout point ambiant à distance au plus r de M possède un unique plus proche voisin sur M. Le reach est un paramètre de convexité généralisé, au sens où M est convexe si et seulement si $\tau_M = \infty$. C'est une notion purement métrique qui permet de quantifier la régularité d'un ensemble sans faire appel à un système de coordonnées particulier. Le fait qu'un ensemble M ait un reach minoré par une constante fixe $\tau_M \geq \tau_{min} > 0$ nous informe à la fois sur ses propriétés locales et globales. En effet, M ne peut alors pas être trop courbée, car τ_{min} prescrit un rayon de courbure minimal pour M. De manière équivalente, on voit que M possède une courbure controllée par $1/\tau_{min}$,

[1] On pourrait traduire *reach* par *la portée*. L'usage a cependant consacré l'emploi de l'anglais.

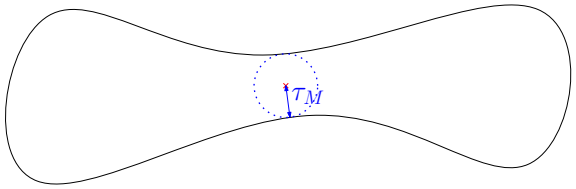


FIGURE *I.1* – Le reach τ_M d'une courbe fermée plane M.

ce que l'on peut assimiler à la classe de Hölder $\mathcal{C}^2(1/\tau_{min})$ en termes de paramétrisations locales. De plus, $\tau_{min} > 0$ empêche M de présenter des régions où elle est proche de s'auto-intersecter (voir Figure I.1), c'est-à-dire des zones d'étranglement arbitrairement petites.

Pour des notions de régularité impliquant des ordres de différentiabilité supérieurs $k \geq 3$ qui s'apparenteraient aux classes $\mathcal{C}^k(L)$, on trouve quelques tentatives de définitions dans la littérature [CP05], mais pas d'étude minimax sur de tels modèles. Pourtant, il semble naturel de pouvoir gagner en vitesse d'estimation lorsque l'objet sous-jacent est plus lisse.

Contributions de cette thèse

Ce manuscrit réunit des résultats d'inférence géométrique issus de trois articles distincts. Les chapitres IV et VI sont le fruit d'une collaboration avec Clément Levrard. Le chapitre V rapporte des travaux effectués avec Jisu Kim, en collaboration avec Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo et Larry Wasserman.

Dans chaque cas, on échantillonne de manière indépendante et identiquement distribuée selon une mesure P ayant pour support M. L'ensemble sous-jacent M est une sous-variété de $\mathbb{R}^D$ au moins de classe $\mathcal{C}^2$. Nous étudions alors, en fonction de la régularité de M, les vitesses minimax d'estimation de fonctionnelles de M. Les fonctionnelles en question sont M elle-même, le reach τ_M, l'espace tangent $T_X M$ et la seconde forme fondamentale II_X^M, pour $X \in M$ à la fois déterministe et aléatoire. La thèse est présentée par ordre croissant de régularité des sous-variétés M étudiées.

Chaque chapitre peut être lu de manière indépendante, ce qui induit quelques redondances, notamment pour ce qui est de la définition des objets. Ainsi, chaque chapitre possède une introduction propre décrivant l'état de l'art pour chaque question traitée. Pour faciliter la présentation des résultats, les preuves et lemmes techniques ont été placés en appendice. Nous détaillons maintenant l'organisation du manuscrit et les principaux résultats obtenus.

Résultats préliminaires

Nous commençons par une partie préliminaire qui introduit des notions courament utilisées en inférence géométrique. Nous établissons quelques résultats techniques à la fois géométriques et probabilistes qui seront utiles par la suite.

Stabilité et optimalité minimax des complexes de Delaunay tangentiels pour la reconstruction de variétés

L'estimation de support, aussi appelée reconstruction de variété en géométrie algorithmique, consiste en l'estimation de M à partir d'un nuage de point $\mathbb{X}_n$ tiré sur ou près de M. Sous une hypothèse de reach analogue à $\mathcal{C}^2(1/\tau_{min})$, il a été prouvé que le complexe de Delaunay tangentiel [BG14] était consistant en l'absence de bruit et lorsque les espaces tangents sont

connus. Plus précisément, si $\mathbb{X}_n$ est assez densément réparti à l'échelle $\varepsilon \leq c_d \tau_M$ sur M, alors le complexe de Delaunay tangentiel est une triangulation ayant pour sommets $\mathbb{X}_n$ qui possède la même topologie que M et qui en est proche à ε^2 pour la distance de Hausdorff d_H. Celui-ci est fournit avec un algorithme polynomial en n permettant de le calculer. Indépendamment, sous la même hypothèse de régularité, les auteurs de [GPPVW12a] ont donné, dans un cadre aléatoire, les vitesses de convergence minimax d'estimation de M pour la distance de Hausdorff. Les auteurs montrent que la vitesse minimax est de l'ordre de $(\log n/n)^{2/d}$. Bien qu'optimal, l'estimateur proposé n'est pas calculable.

Dans ce chapitre, nous démontrons que le complexe de Delaunay tangentiel de [BG14], ajouté à une procédure d'estimation d'espaces tangents basée sur des Analyses en Composantes Principales (ACP) locales, fournit la vitesse optimale $(\log n/n)^{2/d}$ d'estimation pour la distance de Hausdorff donnée dans [GPPVW12a]. Ce résultat reste valide dans un modèle avec bruit additif de petite amplitude.

Inversement, nos résultats montrent que les vitesses optimales [GPPVW12a] sont atteignables avec des triangulations. Ces triangulations sont par ailleurs calculables en temps polynomial.

De plus, en présence de données aberrantes, nous proposons une méthode itérative de débruitage basée sur ces mêmes ACP locales. Le débruitage conduit à la vitesse optimale d'approximation $\big(\log n/(\beta n)\big)^{2/d}$, où $0 < \beta \leq 1$ correspond à la proportion moyenne de points tirés sur M, et $1 - \beta$ celle de données aberrantes. Ici, l'estimation des espaces tangents est utilisée dans la procédure de débruitage et, réciproquement, le nuage de points ainsi débruité permet une estimation plus fine des espaces tangents.

Au cours de l'analyse, on montre que le complexe de Delaunay tangentiel est stable lorsque ses paramètres d'entrée — points et espaces tangents — sont perturbés. L'argument est global, constructif, et peut être appliqué à d'autres méthodes de reconstruction prenant les espaces tangents en paramètre.

Approximation et géométrie du reach

La régularité et les paramètres d'échelle jouent un rôle crucial en analyse de données, en particulier quand il s'agit d'implémentation effective. Comme illustré dans le chapitre IV, le reach τ_M est un paramètre d'échelle et de régularité qui intervient de façon centrale en géométrie algorithmique. Il dicte notamment une échelle minimale pour les caractéristiques géométriques de M.

Nous étudions dans ce chapitre la question de l'estimation du reach à partir de nuages de points $\mathcal{X}$. Avant d'aborder son estimation à proprement parler, on décrit précisément ce à quoi le reach est lié pour les sous-variétés. En particulier, nous démontrons rigoureusement que le reach provient d'une zone de forte courbure (cas local), ou d'une zone d'étranglement (cas global) comme illustré Figure I.1.

Un estimateur plug-in $\hat{\tau}$ de τ_M est proposé lorsque les espaces tangents sont connus et lorsqu'ils sont inconnus. L'analyse des performances de $\hat{\tau}$ est tout d'abord effectuée dans un cadre déterministe, où l'on décrit les propriétés du nuage de points $\mathcal{X}$ qui rendent efficace l'estimation de τ_M par $\hat{\tau}(\mathcal{X})$. Cette analyse est effectuée de manière différente selon que M est dans le cas local ou le cas global, mais l'estimateur $\hat{\tau}$ ne nécessite pas cette information.

Nous examinons l'optimalité de l'estimateur $\hat{\tau}$ via les performances de $\hat{\tau}(\mathbb{X}_n)$ lorsque $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ est un n-échantillon. Lorsque M est dans un modèle $\mathcal{C}^3$, on montre que $\hat{\tau}(\mathbb{X}_n)$ estime τ_M à la vitesse $(1/n)^{2/(3d-1)}$ dans le cas local, et $(1/n)^{1/d}$ dans le cas global. De plus, une borne inférieure sur le risque minimax de l'ordre de $(1/n)^{1/d}$ est obtenue, montrant que $\hat{\tau}(\mathbb{X}_n)$ est optimal dans le cas global.

Dans ce chapitre, le modèle $\mathcal{C}^3$ de régularité est formulé en termes des trajectoires

géodésiques. On suppose que toute géodésique $\gamma(t)$ de M vérifie $\|\gamma'''(t)\| \leq L$. Ici encore, cette notion de régularité ne fait pas appel à un système de coordonnées ambiant.

Vitesses non-asymptotiques d'estimation de variétés, d'espaces tangents et de courbure

Dans ce chapitre, nous étudions les vitesses optimales d'estimation de quantités différentielles associées à des sous-variété jusqu'à l'ordre deux : (0) la sous-variété M elle-même, (1) l'espace tangent $T_X M$ et (2) la seconde forme fondamentale II_X^M, pour $X \in M$ à la fois déterministe et aléatoire.

On introduit une collection de modèles pour les sous-variétés $\mathcal{C}^k$ ($k \geq 3$), qui généralisent de manière naturelle le modèle utilisé au chapitre IV pour $k = 2$. La régularité est exprimée en termes du reach et de l'existence de paramétrisations unitaires bornées dans $\mathcal{C}^k(L)$. On insiste sur la nécessité d'imposer des contraintes à la fois locales et globales pour l'estimation de la courbure et des espaces tangents. En effet, on montre qu'il est impossible d'estimer $T_X M$ et II_X^M si le reach τ_M est autorisé à être arbitrairement petit, malgré des bornes fixes sur la régularité des paramétrisations locales de M.

Les estimateurs proposés sont tous basés sur une unique approche par polynômes locaux, qui généralise l'ACP locale du chapitre IV. On traite ainsi les trois problèmes d'estimation de manière unifiée. À des facteurs $\log n$ près, on montre que les vitesses minimax sont (0) $(\log n/n)^{k/d}$ pour l'estimation de M avec la perte donnée par la distance de Hausdorff, (1) $(\log n/n)^{(k-1)/d}$ pour les espaces tangents et (2) $(\log n/n)^{(k-2)/d}$ pour la seconde forme fondamentale. Autrement dit, la vitesse obtenue pour l'estimation d'une quantité différentielle d'ordre $i = 0, 1, 2$ est $(\log n/n)^{(k-i)/d}$. Les vitesses obtenues montrent en particulier que l'estimation d'espaces tangents par ACP locale du chapitre IV est optimale pour le cas $k = 2$.

Les bornes inférieures minimax sont obtenues via des techniques Bayésiennes déjà connues, bien qu'une nouvelle version conditionnelle du lemme d'Assouad est utilisée pour les espaces tangents et la courbure lorsque le point de base X est aléatoire. En effet, dans ces cas, le paramètre d'intérêt est $T_{X_1} M$ ou $II_{X_1}^M$ avec $\mathbb{X}_n = \{X_1, \ldots, X_n\}$, et l'on est amené à considérer le risque minimax

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} d(\theta_{X_1}(P), \hat{\theta}) = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{n-1}} \left\| d(\theta_{x_1}(P), \hat{\theta}) \right\|_{L^1(P(dx_1))}.$$

Cette écriture montre que la perte considérée est de type L^1, mais la mesure d'intégration P de cette norme L^1 est la loi sous-jacente de l'échantillon, rendant inopérantes les techniques déjà existantes.

Chapter II

General Introduction

As a result of massive data acquisition, the search for relevant features associated to data raises an increasing interest. These features aim at summarize unstructured data, often represented as point clouds in high dimension, by reducing them to descriptors that are simple to analyze. To understand data living in high dimension, a reasonable statistical framework consists in assuming it concentrates on a set of intrinsic dimension d, small compared to the dimension D of the space of measures. This assumption is based on the idea that data are subject to redundancy or correlation, and do not actually include D degrees of freedom.

Linear dimension reduction techniques have been studied extensively. In particular, interests in sparse LASSO-type methods, which aim at setting coefficients of an estimated parameter to zero, have grown considerably [HTF09]. Beyond the developed tools, let us note that the idea of setting coefficients to zero is implicitly based on the reliance on a particular parametrization of the problem. Indeed, the sparsity property of a vector is not stable under deformations, even rigid ones, of the ambient space. When a coordinate system is not reliable, interpretable, or even not available — for instance, distance matrix data [GG12] —, such methods do not apply immediately. In an even more critical situation a regression model may apply in no coordinate system. Indeed, folded up datasets exhibiting nontrivial topology — different from that of a convex — are poorly modeled by graphs of functions. This is the case, for instance, for admissible configurations of some thermodynamic systems, biomolecule conformations, or filamentary distribution of galaxies [LV07]. Then, data present some geometry of which features may be informative, and hence interesting to study.

In a very different context, computational geometry is devoted to study algorithmic problems related to geometry [BY98]. For instance, given a discrete scan of a continuous object, we can try to reconstruct it with triangulations. In this field, theoretical guarantees rely on deterministic sampling conditions, often based on the density and the genericity of a point cloud with respect to the underlying shape. In dimension 2 and 3, geometric feature estimation has been examined widely. The many existing techniques combined with heuristics and efficient data structures provide a satisfactory context for their use [BT07]. In higher dimensions, the literature is quite abundant for topological feature inference, but much less for geometric quantities.

Geometry of Data

Topological data analysis aims at extracting information of geometric or topological nature from data [Car09]. Measures are considered as being generated on "a shape M", which opens many questions on their geometry. This field gets interested in notions that are invariant

under coordinate transformations. De facto, it leads to consider topological invariants and intrinsic quantities coming from differential geometry as providing summaries of data. We can then take advantage of these geometric signatures by using classical learning techniques, such as segmentation or classification [HTF09]. Topological data analysis is at the interface between computational geometry and statistics. It raises the issue of high dimension in computational geometry and, reciprocally, it brings up new descriptors in statistics.

Let us now describe few geometric and topological objects of interest, together with the interpretation on data we can have of it. We take the case considered throughout this thesis, where data come from a source submanifold $M \subset \mathbb{R}^D$ of class at least $\mathcal{C}^2$.

First, the submanifold M itself informs us about location of data. Its estimation is referred to as support, or set estimation [Cue09]. The approximation quality of M by an estimator $\hat{M}$ can be appraised with different losses, depending on the sought properties of approximation. Among the most common ones, let us mention the measure of the symmetric difference $\mu(M \triangle \hat{M})$, that can guarantee fine outlier detection in system monitoring [BCP08]. On the other hand, the Hausdorff distance $d_H(M, \hat{M})$ provides a measure which is rigid enough to guarantee geometric stability properties in regular cases [CCSL06]. Furthermore, the intrinsic dimension $d = \dim(M)$ informs on the degrees of freedom of the underlying system [LV07]. Conversely, the codimension $D - d$ specifies the number of local correlations between variables. In homology, the Betti number $\beta_0(M)$ corresponds to the number of connected components of M. Its preliminary estimation can be necessary in some clustering algorithms [SJ03]. The tangent space T_xM is the best linear approximation of M at $x \in M$. As a consequence, it provides local directions of high variability [ACLZ17]. It is also a good candidate for a domain of local parametrization of dimension d nearby x. As to differential quantities of order two, the second fundamental form II_x^M describes how much M deviates from the linear framework T_xM, as well as the directions where it comes about. It fully encodes curvature, and informs on a local scale at which to look at data [CP05]. Let us also mention more elaborated and global objects such as reach [DS06], volume [BH98], higher order Betti numbers $\beta_k(M)$ [BRS$^+$12], topological persistence [Oud15], Reeb graphs [GSBW11], boundary [CRC04], geodesic distance [MS05] or distance to a measure [CCSM11]. This list is far from exhaustive [Was]. Each of these objects can be a parameter of interest, attached to data, that may be a target for estimation.

This thesis focuses on optimal estimation, from point clouds $\mathbb{X}_n = \{X_1, \dots, X_n\}$, of geometric quantities associated to submanifolds M of the Euclidean space $\mathbb{R}^D$ in a non-asymptotic statistical framework. We investigate optimal rates for estimation of these quantities with different regularity classes for the unknown source submanifold M.

Optimality and Convergence Rates

Up to recently, optimality questions has garnered few attention in geometric inference. In computational geometry, optimality often refers to algorithmic complexity when the problem is of combinatorial nature [AL13]. When it is not the case, the notion of optimality relies on ad hoc constructions of point clouds that are not generic [Cla06]. On the contrary, there are plenty of notions of optimality in statistics, both parametric [LC98] and nonparametric [Tsy09]. In this thesis, we will use the minimax risk, a broadly used optimality criterion in nonparametric statistics. Let us describe its construction in a general framework.

Let us consider the estimation of a parameter of interest $\theta(P)$ depending on the common distribution $P \in \mathcal{P}$ of an n-sample $\mathbb{X}_n = \{X_1, \dots, X_n\}$ with support $M = Supp(P)$. For $\theta(P)$, one can think of the geometric and topological quantities described previously, or more classically, of a regression or density function. We aim at answering the question

"can we estimate $\theta(P)$ from an n-sample $\mathbb{X}_n = \{X_1, \dots, X_n\}$ with distribution P?". By estimating, we mean to find $\hat{\theta} = \hat{\theta}(\mathbb{X}_n)$ making a measure of quality $d(\theta(P), \hat{\theta})$, fixed beforehand, small on average. With P fixed, the Glivenko–Cantelli theorem states that the empirical distribution P_n converges almost surely to P when n goes to infinity. Provided that the functional of interest $P \mapsto \theta(P)$ is stable with respect to P, one can hope to estimate $\theta(P)$, at least asymptotically. However, it is impossible to derive a rate of convergence as n goes to infinity if P is allowed to get close to pathological cases. For instance, with no regularity assumption on a regression function, distinguishing the signal from the noise is hopeless.

To get an answer to the — more precise — question *"How fast can we estimate $\theta(P)$?"*, it is hence necessary to restrict the field of study — a model $\mathcal{P}$ known to contain P — and to examine its intrinsic limitations. The minimax risk $R_n(\mathcal{P})$ on the model $\mathcal{P}$ for the estimation of the parameter $\theta(P)$ with a sample of size n is the best integrated risk that is achievable uniformly on $\mathcal{P}$ by an estimator. Namely,

$$R_n(\mathcal{P}) = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} d(\theta(P), \hat{\theta}),$$

where the infimum ranges among the set of estimators $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$. The minimax risk corresponds to the best attainable performance with n sample points. When n becomes large, $R_n(\mathcal{P})$ informs on the optimal rate of approximation of the quantity of interest $\theta(P)$ on $\mathcal{P}$. Hence, an estimator $\hat{\theta}$ is said to be minimax optimal if, for n large enough,

$$R_n(\mathcal{P}) \leq \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} d(\theta(P), \hat{\theta}) \leq C_{\mathcal{P}} R_n(\mathcal{P}), \tag{II.1}$$

for $C_{\mathcal{P}} > 1$. For studying the behavior of a minimax risk and deriving a result like (II.1), we usually proceed in two very independent steps.

(i) Bounding from above the minimax risk boils down to exhibit an estimator $\hat{\theta}$ and to study its performance uniformly on $\mathcal{P}$. It can be summarized by a bound $\sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} d(\theta(P), \hat{\theta}) \leq v_n$. This bound is problem-specific and relies on nonasymptotic properties of the considered estimator $\hat{\theta}$.

(ii) To show that one cannot achieve better, a lower bound $R_n(\mathcal{P}) \geq v'_n$ is derived with Bayesian arguments [Yu97]. It corresponds to a worst-case study. Indeed, if (at least) two distributions $P_1, P_2 \in \mathcal{P}$ are such that their respective n-samples have distributions that are close, but with parameters $\theta(P_1)$ and $\theta(P_2)$ far away from each other, then no estimator can be both close to $\theta(P_1)$ and $\theta(P_2)$ simultaneously. Hence, no estimator can be accurate at a scale smaller than $d(\theta(P_1), \theta(P_2))$ with high probability.

If $v_n \leq C_{\mathcal{P}} v'_n$, we get (II.1) and we say that v_n, or equivalently v'_n, is the optimal rate of estimation of $\theta(P)$ on $\mathcal{P}$.

Here, considering a mean risk on n-samples allows to formalize the notion of genericity of a point cloud. Therefore, a procedure that outputs a quantity $\hat{\theta}(\mathcal{X})$ built on top of n point $\mathcal{X} = \{x_1, \dots, x_n\}$ (seen as deterministic) may be considered as generically optimal if it reaches the minimax rate when evaluated on a n-sample $\mathbb{X}_n$. Randomness provides a framework where the notion of optimality is well-posed.

Quantitative Notions of Geometric Regularity

As described previously, a minimax study requires to specify a model $\mathcal{P}$ beforehand. It defines the regularity class of the objects in question. Here, regularity is to be understood

in the broad sense, and may include dimension, space massiveness, approximability of objects, or smoothness in the classical differential sense. The broader this class is, the more general and difficult the problem becomes. The smaller it is, the more specific and accessible the estimation is. Defining a model boils down to characterize quantitatively the difficulty of estimation of an object.

By quantitative regularity, we express the need for bounding the studied objects. In analogy with regression, classes $\mathcal{C}^k$ of functions k times continuously differentiable do not furnish a sufficient information to exploit uniformly their smoothness. Density of $\mathcal{C}^k$ in the set of continuous maps shows that it is possible to get close to non-smooth pathological cases. On the contrary, Hölder classes $\mathcal{C}^k(L)$ — composed of $k-1$ times differentiable functions with L-Lipschitz $(k-1)$th derivative — avoid this phenomenon. Similarly, for models of submanifolds $M \subset \mathbb{R}^D$, smoothness must be quantified. Yet, the absence of a canonical coordinate system makes the subject more intricate than for functions.

For a submanifold M, a first regularity characteristic is its intrinsic dimension $d = \dim(M)$. It drives its metric massiveness in the first place as in the Euclidean case. Thereafter, we will always assume the dimension d to be known. We will see that d impacts considerably the rates of estimation of the objects studied. Besides, we will carefully develop an analysis that is unaffected by the ambient dimension D which, extrinsic, may be very large compared to d.

Then, for fixed dimension d, a very popular regularity parameter in geometric inference is the reach. First introduced by Herbert Federer for geometric measure theory [Fed59], the reach τ_M of $M \subset \mathbb{R}^D$ is the largest radius $r \geq 0$ such that any ambient point at distance at most r from M has a unique nearest neighbor on M. The reach is a generalized convexity

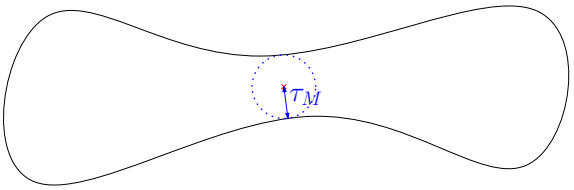


Figure II.1 – The reach τ_M of a closed curve M in the plane.

parameter, in the sense that M is convex if and only if $\tau_M = \infty$. It is a purely metric notion that allows to quantify the regularity of a set without regard of any particular coordinate system. For a set M, having its reach bounded from below by a fixed constant $\tau_M \geq \tau_{min} > 0$ informs us both on its local and global properties. Indeed, M cannot be too curved, since τ_{min} prescribes a minimal radius of curvature for M. Equivalently, we see that M has curvature controlled by $1/\tau_{min}$, which we can think of as a Hölder class $\mathcal{C}^2(1/\tau_{min})$ in terms of local parametrizations. Furthermore, $\tau_{min} > 0$ prevents M to contain regions where it is close to self-intersect. That is, arbitrarily narrow bottleneck structures (see Figure II.1).

For regularity notions involving higher orders of differentiability $k \geq 3$, which would be similar to $\mathcal{C}^k(L)$ classes, one can find some attempts of definitions in the litterature [CP05], but no minimax study on such models. Though, it seems natural to get faster estimation rates when the underlying object is smoother.

Contribution of this Thesis

This thesis manuscript gathers results on geometric inference coming from three distinct articles. Chapters IV and VI are the results of a collaboration with Clément Levrard.

Chapter V relates works carried out with Jisu Kim, in collaboration with Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo and Larry Wasserman.

In each case, we sample independently a distribution P having support M. The underlying set M is a submanifold of $\mathbb{R}^D$ of class at least $\mathcal{C}^2$. We study, depending on the regularity of M, the minmax rates of estimation of functionals of M. The studied functionals are M itself, the reach τ_M, the tangent space $T_X M$ and the second fundamental form II_X^M, for $X \in M$ deterministic and random. The thesis is exposed in ascending order of regularity of the submanifolds M.

Each chapter can be read independently, which gives rise to few redundancies, especially for definitions of objects. Thus, each chapter has its own introduction describing the state of the art for each question we deal with. For ease of exposition, the proofs and technical lemmas are placed in appendices. Let us now detail the overall organization of the thesis, together with the main results we obtained.

Preliminary Results

We get started with a preliminary chapter in which we introduce notions that are commonly used in geometric inference. We derive both geometric and probabilistic technical results that will be useful later on.

Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction

Support estimation, also referred to as manifold reconstruction in computational geometry, consists in the estimation of M from point cloud — say $\mathbb{X}_n$ — drawn on or nearby M. Under a reach regularity condition similar to $\mathcal{C}^2(1/\tau_{min})$, the tangential Delaunay complex [BG14] was shown to be consistent if no noise is present and tangent spaces are known. More precisely, if $\mathbb{X}_n$ is distributed densely enough at scale $\varepsilon \leq c_d \varepsilon_0$ on M, then the tangential Delaunay complex is a triangulation with vertices $\mathbb{X}_n$ that has the same topology as M, and is ε^2-close to it for the Hausdorff distance. It is computable with an algorithm which is polynomial in n. Independently, under the same regularity assumptions, the authors of [GPPVW12a] derived, in a random framework, minimax rates of estimation of M for the Hausdorff distance. The authors showed that the minimax rate is of order $(\log n/n)^{2/d}$. Though optimal, the estimator of [GPPVW12a] is not constructive.

In this chapter, we show that the tangential Delaunay complex of [BG14], together with a tangent space estimation procedure based on local Principal Component Analysis (PCA), yields the optimal estimation rate $(\log n/n)^{2/d}$ for the Hausdorff distance that was given in [GPPVW12a]. This result still holds in a model with additive noise of small amplitude.

Conversely, our results show that the optimal rates [GPPVW12a] are achievable with triangulations. In addition, these triangulations are computable in polynomial time.

Furthermore, in the presence of outliers, we propose an iterative denoising method based on the same local PCA's. Denoising leads up to the optimal approximation rate $(\log n/(\beta n))^{2/d}$, where $0 < \beta \leq 1$ is the average proportion of points drawn on M, and $1-\beta$ that of outliers. Here, tangent space estimation is used in the denoising procedure and, conversely, the denoised point cloud allows for a more accurate tangent space estimation.

In the analysis, we show that the tangential Delaunay complex is stable when its input parameters — points and tangent spaces — are perturbed. The argument is global, constructive, and may be applied to other reconstruction methods taking tangent spaces as input.

Approximation and Geometry of the Reach

Regularity and scale parameters play a crucial role in data analysis, in particular for effective implementation. As illustrated in Chapter IV, the reach τ_M is one that appears in a central way in computational geometry. In particular, it gives a minimal scale of geometric features of M.

In this chapter, we study the problem of estimation of the reach from a point cloud $\mathcal{X}$. Before tackling estimation itself, we describe precisely what reach is linked to for submanifolds. In particular, we show in a rigorous way that the reach either comes from a highly curved zone (local case), or from a bottleneck structure (global case) as illustrated Figure II.1.

A plugin estimator $\hat{\tau}(\mathcal{X})$ of τ_M is proposed when tangent spaces are both known and unknown. The analysis of performances of $\hat{\tau}$ are first carried out in a deterministic framework, by describing the properties of $\mathcal{X}$ making efficient the estimation of τ_M by $\hat{\tau}(\mathcal{X})$. This analysis is different depending on whether M is in the local or the global case, but the estimator $\hat{\tau}$ does need this information.

We advocate the optimality of $\hat{\tau}$ with the performances of $\hat{\tau}(\mathbb{X}_n)$, when $\mathbb{X}_n$ is an n-sample. When M belongs to a $\mathcal{C}^3$ model, we show that $\hat{\tau}(\mathbb{X}_n)$ approximates τ_M at rate $(1/n)^{2/(3d-1)}$ in the local case, and $(1/n)^{1/d}$ in the global case. Moreover, a minimax lower bound of order $(1/n)^{1/d}$ is derived, showing that $\hat{\tau}(\mathbb{X}_n)$ is optimal in the global case.

In this chapter, the $\mathcal{C}^3$ model of regularity is expressed in terms of geodesic trajectories. We assume that any geodesic $\gamma(t)$ of M satisfies $\|\gamma'''(t)\| \leq L$. Here, again, this regularity notion does not rely on an ambient coordinate system.

Non-Asymptotic Rates for Manifold, Tangent Space and Curvature Estimation

In this chapter, we study the optimal rates of estimation for differential quantities associated to submanifolds up to order two: (0) the submanifold M itself, (1) the tangent space $T_X M$ and (2) the second fundamental form II_X^M, for $X \in M$ both deterministic and random.

We introduce a collection of models for $\mathcal{C}^k$-submanifolds ($k \geq 3$), that generalize naturally the model used in Chapter IV for $k = 2$. Regularity is expressed in terms of the reach and of the existence of unit parametrizations that are bounded in $\mathcal{C}^k(L)$. We insist on the need to impose both local and global constraints for estimation of curvature and tangent spaces. Indeed, we show that it is impossible to estimate $T_X M$ and II_X^M if the reach τ_M is allowed to be arbitrarily small, despite fixed bounds on the regularity of local parametrizations of M.

The proposed estimators are all based on a single approach of local polynomial fitting. It generalizes local PCA of Chapter IV. Thus, we deal with the three estimation problems in a unified way. Up to $\log n$ terms, we show that the minimax rates are (0) $(\log n/n)^{k/d}$ for the estimation of M with the loss given by the Hausdorff distance, (1) $(\log n/n)^{(k-1)/d}$ for tangent spaces, and (2) $(\log n/n)^{(k-2)/d}$ for the second fundamental form. In other words, the rate for a differential quantity of order i is $(\log n/n)^{(k-i)/d}$. In particular, the derived rates show that the tangent space procedure of Chapter IV, based on local PCA, is optimal for $k = 2$.

Minimax lower bounds are derived with existing Bayesian techniques, although a new conditional version of Assouad's lemma is used for tangent spaces and curvature when the base point X is random. Indeed, in this case, the parameter of interest is $T_{X_1} M$ or $II_{X_1}^M$,

with $\mathbb{X}_n = \{X_1, \ldots, X_n\}$, which leads to consider the minimax risk

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} d(\theta_{X_1}(P), \hat{\theta}) = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{n-1}} \left\| d(\theta_{x_1}(P), \hat{\theta}) \right\|_{L^1(P(dx_1))} .$$

This formulation shows that the loss involved is of type L^1. However, the integration measure P of this L^1-norm is the underlying distribution of the sample, which makes the existing techniques unsuccessful.

Chapter III

Preliminary Results

Abstract

We introduce notions that are commonly used in geometric inference, such as the Hausdorff distance and the reach. We derive both geometric and probabilistic technical results that will be useful later on.

Content

III.1 Hausdorff Distance and Measurability

In Chapters IV and VI, we tackle reconstruction problems. Given random variables $X_1, \dots, X_n$, we build — with various methods depending on the context — estimators $\hat{M} = \hat{M}(X_1, \dots, X_n)$ aiming at approximate a target compact subset $M \subset \mathbb{R}^D$. Hence, we have to make clear what "approximate" means for compact sets. For this, we use the Hausdorff distance d_H. Consequently, we have to clarify what to be a compact sets-valued "estimator" means, or equivalently, describe measurability properties in the space of compact subsets endowed with the Hausdorff distance. To ensure not to focus on technical details about measurability later on, we choose to address this in this section.

Roughly speaking, the take-away message is that the class of compact subsets of a metric space behaves as well as the metric space itself. Hence, random variables with values in it do so. For (much) more details about measurability in classes of subspaces, we refer to [Mat75], and to [Bee93] for the functional approach we adopt.

III.1.1 Hausdorff Distance

Let $(\mathcal{D}, d)$ be a metric space. This thesis only tackles the case $(\mathcal{D}, d) = (\mathbb{R}^D, \|\cdot\|)$. However, we state Hausdorff distance properties in full generality to emphasize the key points that have our case work. We let $\mathcal{K}(\mathcal{D})$ denote the set of nonempty compact subsets of $(\mathcal{D}, d)$. For $x \in \mathcal{D}$ and $K \subset \mathcal{D}$, the distance from x to K is

$$d(x, K) = \inf \{d(x, y), y \in K\}. \tag{III.1}$$

One easily checks that $d(., K)$ is a 1-Lipschitz map. Let us define the Hausdorff distance.

Definition III.2. *For two compact subsets $A, B \subset \mathbb{R}^D$, the Hausdorff distance between A and B is defined by*

$$d_H(A, B) = \max\{\sup_{a \in A} d(a, B), \sup_{b \in B} d(b, A)\}.$$

d_H is a distance on the space $\mathcal{K}(\mathcal{D})$ of nonempty compact subsets of $(\mathcal{D}, d)$.

An equivalent formulation of d_H can be written in terms of offsets. The r-offset of K is

$$K^r = \{x \in \mathcal{D}, d(x, K) \leq r\}, \tag{III.3}$$

the set of ambient points that are at distance less than or equal to r from K.

Proposition III.4. *For all $A, B \in \mathcal{K}(\mathcal{D})$,*

$$d_H(A, B) = \inf \{r > 0, A^r \supset B \text{ and } B^r \supset A\}.$$

The Hausdorff distance is a rigid distance, in the sense a single point added to a set — say, an outlier — can have the Hausdorff distance blow up, since $d_H(A, A \cup x) = d(x, A)$. It plays the role of a L^∞ dissimilarity in the space of compact sets. One can make this idea precise by identifying a compact subset $A \subset \mathcal{D}$ to its distance function $d(\cdot, A)$, which is locally bounded.

Proposition III.5. *The map*

$$\begin{aligned}(\mathcal{K}(\mathcal{D}), d_H) &\longrightarrow (\mathcal{C}(\mathcal{D}, \mathbb{R}_+), \|\cdot\|_\infty) \\ A &\longmapsto d(\cdot, A)\end{aligned}$$

is an isometry. In other words, for all $A, B \in \mathcal{K}(\mathcal{D})$,

$$d_H(A, B) = \sup_{x \in \mathcal{D}} |d(x, A) - d(x, B)|.$$

Moreover, if $(\mathcal{D}, d)$ is complete, $(\mathcal{K}(\mathcal{D}), d_H)$ is closed in $(\mathcal{C}(\mathcal{D}, \mathbb{R}_+), \|\cdot\|_\infty)$.

Actually, we prove the slightly stronger identity

$$d_H(A, B) = \sup_{x \in K} |d(x, A) - d(x, B)|,$$

for all $A \cup B \subset K \subset \mathcal{D}$, meaning that one can restrict the distance function to the domain $A \cup B$ to compare A and B. When measuring the dissimilarity between compact subsets, we can somehow restrict to the geometry of $(A \cup B, d)$.

By identifying a compact subset with its associated distance function, one can see $(\mathcal{K}(\mathcal{D}), d_H)$ as a closed subset of $(\mathcal{C}(\mathcal{D}, \mathbb{R}_+), \|\cdot\|_\infty)$. Consequently, it inherits its usual topological and metric properties. Conversely, one has the isometric closed inclusion

$$\begin{aligned}(\mathcal{D}, d) &\longrightarrow (\mathcal{K}(\mathcal{D}), d_H)\\ x &\longmapsto \{x\},\end{aligned}$$

that allows to identify a point x to the singleton $\{x\}$. Hence, roughly speaking, $(\mathcal{K}(\mathcal{D}), d_H)$ cannot have better metric properties than $(\mathcal{D}, d)$. We recall that a metric space is said to be boundedly compact if all its closed bounded subsets are compact. In particular, a boundedly compact metric space is complete.

Proposition III.6. *Let $(\mathcal{D}, d)$ be a metric space.*

(i) $(\mathcal{D}, d)$ is separable if and only if $(\mathcal{K}(\mathcal{D}), d_H)$ is separable,

(ii) $(\mathcal{D}, d)$ is compact if and only if $(\mathcal{K}(\mathcal{D}), d_H)$ is compact,

(iii) $(\mathcal{D}, d)$ is boundedly compact if and only if $(\mathcal{K}(\mathcal{D}), d_H)$ is boundedly compact,

(iv) $(\mathcal{D}, d)$ is complete if and only if $(\mathcal{K}(\mathcal{D}), d_H)$ is complete,

(v) $(\mathcal{D}, d)$ is Polish if and only if $(\mathcal{K}(\mathcal{D}), d_H)$ is Polish.

To avoid measure-theoretic difficulties, the mildest framework commonly adopted to develop probability theory is random variables with values in Polish spaces [Par05]. Hence, working in $(\mathcal{K}(\mathcal{D}), d_H)$ when $(\mathcal{D}, d)$ is Polish will have all the usual probability theory tools operate in a non-pathological way. In particular, manipulating random variables in $(\mathcal{K}(\mathbb{R}^{\mathcal{D}}), d_H)$ will not raise any specific issue.

III.1.2 Compact Set-Valued Random Variables

Now that we made sure handling compact sets-valued random variables is not problematic, let us describe some of them in the case $(\mathcal{D}, d) = (\mathbb{R}^D, \|\cdot\|)$. We give a few examples of measurable maps in $\mathcal{K}(\mathbb{R}^D)$ endowed with the Borel σ-field associated to the Hausdorff metric d_H. We let $\mathcal{C}(\mathbb{R}^D, \mathbb{R}^D)$ denote the set of continuous map from $\mathbb{R}^D$ to itself, that we endow with the topology of the uniform convergence on compact sets, and its Borel σ-field.

Proposition III.7. *Equip $\mathcal{K}(\mathbb{R}^D)$ with the Borel σ-field associated to the Hausdorff metric d_H. Then the following maps are measurable:*

(i) $\mathbb{R}^D \ni x \longmapsto \{x\}$, for all $x \in \mathbb{R}^D$,

(ii) $\mathcal{C}(\mathbb{R}^D, \mathbb{R}^D) \times \mathcal{K}(\mathbb{R}^D) \ni (f, A) \mapsto f(A) = \{f(x), x \in A\}$,

(iii) $\mathcal{K}(\mathbb{R}^D) \times \mathcal{K}(\mathbb{R}^D) \ni (A, B) \longmapsto A \cup B$,

(iv) $\mathcal{K}(\mathbb{R}^D) \ni A \mapsto conv(A)$.

By composition, Proposition III.7 actually allows to describe a wide variety of measurable maps. For instance, a simplicial complex is a finite union of simplicies, and simplicies are convex hulls of finite sets. As a consequence, (i),(ii) and (iv) show that simplicial complexes $\hat{M}$ built on top of a random point cloud $\mathbb{X}_n$ for which the presence of each simplex is determined by a measurable event, yield estimators. As a consequence, the simplicial complexes $\hat{M}_{\texttt{TDC}}$, $\hat{M}_{\texttt{TDC}\delta}$ and $\hat{M}_{\texttt{TDC}+}$ of Chapter V are measurable. Similarly, the union of local polynomial patches $\hat{M}_{\texttt{POLY}}$ of Chapter VI are measurable from (ii) and (iii).

III.2 Measure, Diameter, and Sampling

In this section, we consider a general metric-measure space $(\mathcal{D}, d, \mu)$ [1] , and we investigate the links between the local behavior of μ and metric properties of $\mathcal{D}$.

For sake of simplicity, $(\mathcal{D}, d)$ will be assumed to be separable. All the metric quantities considered here refer to d: balls $\mathcal{B}(x, r)$, diameter $diam(\cdot)$, Hausdorff distance $d_H(\cdot, \cdot)$, and so on. The support $Supp(\mu) \subset \mathcal{D}$ of μ is the smallest closed set $C \subset \mathcal{D}$ of mass one.

Definition III.8 (Standard Measure)**.** *The distribution μ is called (a, b)-standard at scale r_0 if for all $x \in Supp(\mu)$ and all $r \leq r_0$,*

$$\mu\left(\mathcal{B}(x, r)\right) \geq a r^b.$$

Roughly speaking, a measure that is (a, b)-standard at scale r_0 behaves like the b-dimensional Lebesgue measure, though b need not be an integer. This assumption is pretty popular in the literature on set estimation, and its properties will be used extensively in the results of this thesis. Up to now, it was considered with $b = D$ in $\mathbb{R}^D$ [Cue09], except in [CCSL06] in a theoretical framework. As we will see shortly, such an assumption gives bounds on massiveness of the support $Supp(\mu) \subset \mathcal{D}$.

To measure massiveness of subsets $K \subset \mathcal{D}$, we will use packing and covering numbers. That is, numbers of balls optimally displayed at some scale r in K. A r-covering of K is a subset $\mathbf{x} = \{x_1, \ldots, x_k\} \subset K$ such that for all $x \in K$, $d(x, \mathbf{x}) \leq r$. A r-packing of K is a subset $\mathbf{y} = \{y_1, \ldots, y_k\} \subset K$ such that for all $y, y' \in \mathbf{y}$, $\mathcal{B}(y, r) \cap \mathcal{B}(y', r) = \emptyset$ (or equivalently $d(y, y') > 2r$).

Definition III.9. *For $K \subset \mathcal{D}$ and $r > 0$, the covering number $\mathrm{cv}(K, r)$ is the minimum number of balls of radius r that are necessary to cover K,*

$$\mathrm{cv}(K, r) = \min\left\{k > 0, \text{ there exists a } r\text{-covering of cardinality } k\right\}.$$

The packing number $\mathrm{pk}(K, r)$ is the maximum number of disjoint balls of radius r that can be packed in K,

$$\mathrm{pk}(K, r) = \max\left\{k > 0, \text{ there exists a } r\text{-packing of cardinality } k\right\}.$$

Usually, for a given space K, covering and packing numbers have the same order of magnitude. Namely, they are linked by the relations

$$\mathrm{pk}(K, 2r) \leq \mathrm{cv}(K, 2r) \leq \mathrm{pk}(K, r). \tag{III.10}$$

Indeed, for the left-hand side inequality, notice that if K is covered by a family of balls of radius $2r$, each of these balls contain at most one point of a maximal packing $\mathbf{y}$ at scale $2r$. Conversely, the right-hand side inequality follows from the fact that a maximal r-packing $\mathbf{y}$ is always a $2r$-covering. If it was not the case, one could add a point x_0 such that $d(x_0, \mathbf{y}) > 2r$, which is impossible by maximality of $\mathbf{y}$.

It is interesting to note that any (a, b)-standard measure μ has support massiveness controlled, in the following sense.

Proposition III.11. *Let μ be a (a, b)-standard probability distribution at scale $r_0 > 0$. Then for $r \leq r_0$,*

$$\mathrm{pk}\left(Supp(\mu), r\right) \leq \frac{1}{a r^b}.$$

[1] That is, a metric space $(\mathcal{D}, d)$ with a probability distribution μ on $\mathcal{D}$ equipped with its Borel σ-algebra.

For $r \leq 2r_0$,

$$\mathrm{cv}\left(Supp(\mu), r\right) \leq \frac{2^b}{ar^b}.$$

As a consequence, one can derive an upper bound on the diameter of such a support, if it is assumed to be path-connected. This is based on the following bound.

Lemma III.12. *Let $K \subset \mathbb{R}^D$ be a bounded subset. If K is path-connected, then for all $\varepsilon > 0$, $\mathrm{diam}(K) \leq 2\varepsilon \mathrm{cv}(K, \varepsilon)$.*

Thereby, Proposition III.13 follows from Lemma III.12 applied with $r = 2r_0$, together with Proposition III.11.

Proposition III.13. *If μ is (a, b)-standard at scale r_0 and has a path-connected support $Supp(\mu)$, then*

$$diam\left(Supp(\mu)\right) \leq 4r_0^{1-b}/a.$$

Note that connectedness is crucial here. Consider for instance $K_x = \mathcal{B}(-x, 1) \cup \mathcal{B}(x, 1) \subset \mathbb{R}^D$ for $\|x\|$ arbitrarily large. Then $diam(K_x) = 2\|x\| + 1 \to \infty$ although the distribution on K_x is (a, D)-standard at scale 1 with fixed $a > 0$. Proposition III.13, yields that for all $x \in Supp(\mu)$ (path-connected) and $r > 0$, we have

- if $r \leq r_0$, $\mu\left(\mathcal{B}(x, r)\right) \geq ar^b$,
- if $r \geq diam(Supp(\mu))$, $\mu\left(\mathcal{B}(x, r)\right) = 1$,
- if $r_0 \leq r \leq diam(Supp(\mu))$,

$$\begin{aligned}
\mu\left(\mathcal{B}(x, r)\right) &\geq \mu\left(\mathcal{B}(x, r_0)\right) \\
&\geq ar_0^b = a\left(\frac{r_0}{r}\right)^b r^b \\
&\geq a\left(\frac{ar_0}{4r_0^{1-b}}\right)^b r^b = a\left(\frac{ar_0^b}{4}\right)^b r^b.
\end{aligned}$$

In turn, if $Supp(\mu)$ is path-connected, $\mu(\mathcal{B}(x, r)) \geq a'r^b \wedge 1$ for all $r > 0$, where $a' \triangleq a\big(ar_0^b/4\big)^b$. This property was used in [CCSL06] and called "(a, b)-standardness" (with no scale). Conversely, if $\mu\left(\mathcal{B}(x, r)\right) \geq a'r^b \wedge 1$ for all $r > 0$, then $\mu\big(\mathcal{B}(x, a'^{1/b})\big) = 1$ for all $x \in Supp(\mu)$, so that $diam(Supp(\mu)) \leq a'^{1/b}$.

Further investigating the properties of (a, b)-standard measures at scale $r_0 > 0$, let us now give the convergence rate of a sample point cloud $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ towards its underlying support $Supp(\mu)$.

Proposition III.14. *If μ is (a, b)-standard at scale $r_0 > 0$, and $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ is an i.i.d. n-sample with common distribution μ, then for all $r \leq 2r_0$,*

$$\mathbb{P}\left(d_H\big(Supp(\mu), \mathbb{X}_n\big) > r\right) \leq \frac{4^b}{ar^b} \exp\left(-nar^b\right).$$

In particular, for any $\alpha > 0$, for n large enough so that $\left(C_{a,b,\alpha}\frac{\log n}{n}\right)^{1/b} \leq 2r_0$, with probability at least $1 - \left(\frac{1}{n}\right)^\alpha$,

$$d_H\big(Supp(\mu), \mathbb{X}_n\big) \leq \left(C_{a,b,\alpha}\frac{\log n}{n}\right)^{1/b},$$

where $C_{a,b,\alpha} = \frac{(1+\alpha)\vee 4^b}{a}$.

In other words, with n points, the typical density of sampling of a (a, b)-standard measure is of order $(\log n/n)^{1/b}$. Roughly speaking, it relies on the fact that standard measures have uniformly spread mass on their support. Hence, an n-sample would visit all the areas of its support with high probability. This comes from the fact that the massiveness of $Supp(\mu)$ (in terms of covering number) is controlled.

III.3 Reach and Submanifolds of $\mathbb{R}^D$

We now introduce the reach and describe links between curvature, reach, diameter, and volume in the case of submanifolds.

III.3.1 Reach of Closed Subsets

Let us first describe the reach in full generality, as first introduced by Federer [Fed59]. Given a closed subset $A \subset \mathbb{R}^D$, the medial axis $Med(A)$ of A is the subset of $\mathbb{R}^D$ consisting of points that have at least two nearest neighbors on A. Namely, denoting by $d(z, A) = \inf_{p\in A} \|p - z\|$ the distance function to A,

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

The reach of A is then defined as the minimal distance from A to $Med(A)$.

Definition III.15. *The reach of a closed subset $A \subset \mathbb{R}^D$ is defined as*

$$\tau_A = \inf_{p\in A} d\left(p, Med(A)\right) = \inf_{z\in Med(A)} d\left(z, A\right). \tag{III.16}$$

Some authors refer to τ_A^{-1} as the *condition number* [NSW08, SW12]. Indeed, the value of the reach quantifies the degree of regularity of a set, with larger values associated to more regular sets. From the definition of the medial axis, the projection $\pi_A(x) = \arg\min_{p\in A} \|p - x\|$ onto A is well defined outside $Med(A)$. The reach is the largest distance $\rho \geq 0$ such that π_A is well defined on the ρ-offset $A^\rho = \{x \in \mathbb{R}^D | d(x, A) \leq \rho\}$. Hence, the condition $\tau_A \geq \tau_{min}$ can be seen as a generalization of convexity, since a set $A \subset \mathbb{R}^D$ is convex if and only if $\tau_A = \infty$.

Positive reach is the minimal regularity assumption in geometric measure theory and integral geometry [Fed69]. Sets with positive reach exhibit a structure that is close to be differential, with the so-called tangent and normal cones [Fed59]. It is a $\mathcal{C}^2$ notion, in the sense that it is stable under $\mathcal{C}^2$ ambient deformations. Let us state a stability result for the reach with respect to $\mathcal{C}^2$ diffeomorphisms.

Lemma III.17 (Theorem 4.19 in [Fed59])**.** *Let $A \subset \mathbb{R}^D$ be a closed subset with $\tau_A \geq \tau_{min} > 0$ and $\Phi : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ is a $\mathcal{C}^1$-diffeomorphism such that Φ,Φ^{-1}, and $d\Phi$ are Lipschitz with Lipschitz constants K,N and R respectively, then*

$$\tau_{\Phi(A)} \geq \frac{1}{(K\tau_{min}^{-1} + R)N^2}.$$

We do not detail further properties of sets with positive reach with no extra regularity assumption. The interested reader may refer to the original article [Fed59], and to [Thä08] for a more recent review.

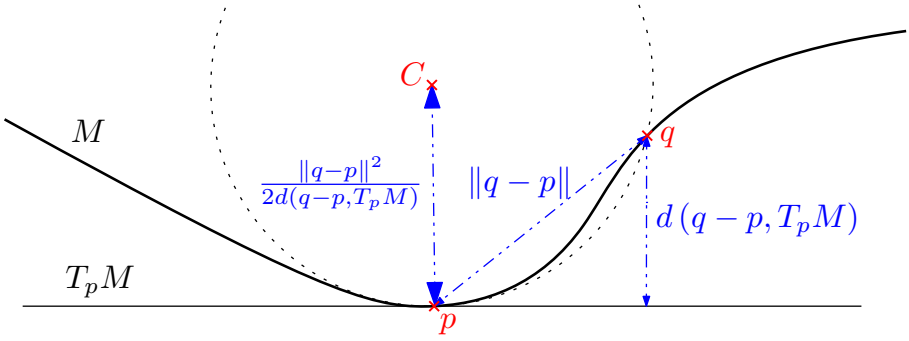


Figure III.1 – Geometric interpretation of quantities involved in (V.5).

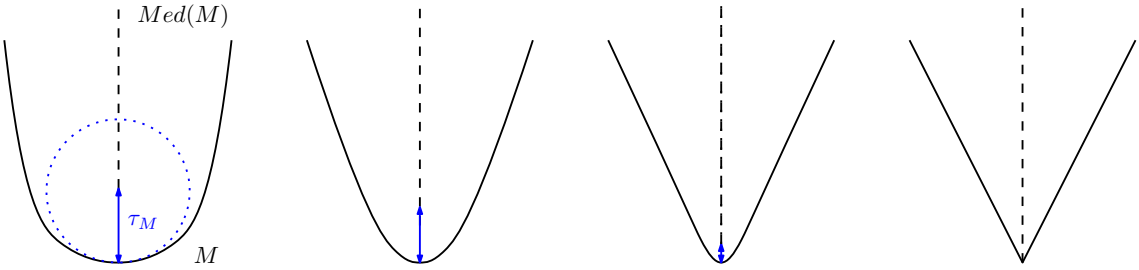


Figure III.2 – Small reach τ_M may witness high curvature.

III.3.2 Geometry of Submanifolds with Reach Bounded Away from Zero

In the case of submanifolds, one can reformulate the definition of the reach in the following manner. Here, for all $p \in M$, T_pM stands for the tangent space of M at p [dC92, Chapter 0].

Theorem III.18 (Theorem 4.18 in [Fed59]). *For any submanifold $M \subset \mathbb{R}^D$,*

$$\tau_M = \inf_{q \neq p \in M} \frac{\|q-p\|^2}{2d(q-p, T_pM)}. \tag{III.19}$$

Another way to state (III.19) is that $\tau_M \geq \tau_{min} > 0$ if and only if for all $p, q \in M$,

$$d(q-p, T_pM) \leq \frac{\|q-p\|^2}{2\tau_{min}}.$$

In other words, one gets a quantitative bound on how fast the submanifold M deviates from its tangent spaces. By definition of tangent spaces, this happens at most with a quadratic growth $\mathcal{O}(\|q-p\|^2)$, but the reach allows for an explicit constant and no restriction of $\|q-p\|$ small. Furthermore, the ratio appearing in (III.19) can be interpreted geometrically, as suggested in Figure III.1. It is the radius of an ambient ball, tangent to M at p and passing through q. Hence, at a differential level, the reach gives an upper bound on the radii of curvature of M. Equivalently, τ_M^{-1} is a bound on the curvature of M, as illustrated Figure III.2.

Proposition III.20 (Proposition 6.1 in [NSW08]). *Let $M \subset \mathbb{R}^D$ be a compact submanifold with reach $\tau_M \geq \tau_{min} > 0$, and γ an arc-length parametrized geodesic of M. Then for all t,*

$$\|\gamma''(t)\| \leq 1/\tau_{min}.$$

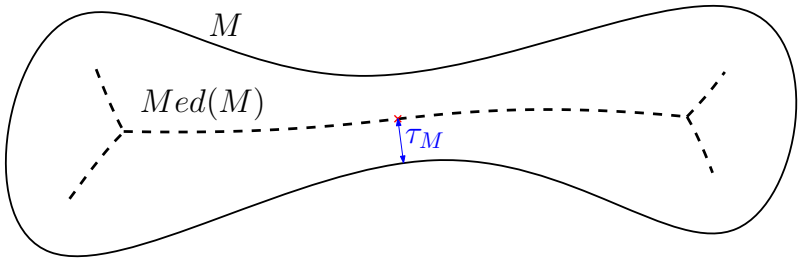


Figure III.3 – Small reach τ_M may witness a narrow bottleneck structure.

Furthermore, as illustrated in Figure III.3, the condition $\tau_M \geq \tau_{min} > 0$ also prevents bottleneck structures where M is nearly self-intersecting.

As a consequence, at a scale of order τ_M, one can link the (extrinsic) Euclidean distance to the (intrinsic) geodesic distance d_M of M. The precise bound is the following.

Lemma III.21 (Proposition 6.3 in [NSW08])**.** *If $\tau_M \geq \tau_{min}$, then for all $p, q \in M$ such that $\|p - q\| \leq \tau_{min}/2$,*

$$\|q - p\| \leq d_M(p, q) \leq \tau_{min} \left(1 - \sqrt{1 - \frac{2\|p - q\|}{\tau_{min}}}\right).$$

Let us move to other bounds on differential geometric quantities. For this, we need notation from differential geometry. First, $\mathcal{B}_M$ denote closed balls for the geodesic distance d_M, and $\mathring{\mathcal{B}}, \mathring{\mathcal{B}}_M$ are open balls. We let $II_p^M : T_pM \times T_pM \to T_pM^\perp$ denote the second fundamental form of M at p [dC92, p. 125]. II_p^M characterizes the curvature of M at p. For all $p \in M$ and all unit $v \in T_pM$, we denote by $\gamma_{p,v}$ the unique arc-length parametrized geodesic of M such that $\gamma_{p,v}(0) = p$ and $\gamma'_{p,v}(0) = v$. The exponential map is then defined as $\exp_p(v) = \gamma_{p,v}(1)$. Finally, $\mathcal{H}^d$ stands for the d-dimensional Hausdorff measure on $\mathbb{R}^D$ [Fed69, p. 171].

Proposition III.22. *Let $M \subset \mathbb{R}^D$ be a d-dimensional submanifold with reach $\tau_M \geq \tau_{min} > 0$.*

(i) For all $p \in M$ and all unit $v \in T_pM$, $\left\|II_p^M(v, v)\right\| \leq 1/\tau_{min}$.

(ii) The injectivity radius of M is at least $\pi\tau_{min}$. That is, for all $p \in M$, the map $\exp_p : \mathring{\mathcal{B}}_{T_pM}(0, \pi\tau_{min}) \to \mathring{\mathcal{B}}_M(p, \pi\tau_{min})$ is a diffeomorphism

(iii) The sectional curvatures κ of M satisfy $-\frac{2}{\tau_{min}^2} \leq \kappa \leq \frac{1}{\tau_{min}^2}$.

(iv) For all $\|v\| < \frac{\pi\tau_{min}}{2\sqrt{2}}$ and $w \in T_pM$,

$$\left(1 - \frac{\|v\|^2}{6\tau_{min}^2}\right)\|w\| \leq \left\|d_v \exp_p \cdot w\right\| \leq \left(1 + \frac{\|v\|^2}{\tau_{min}^2}\right)\|w\|$$

(v) For all $p \in M$, $r \leq \frac{\pi\tau_{min}}{2\sqrt{2}}$, and a Borel set $A \subset \mathcal{B}_{T_qM}(0, r) \subset T_qM$,

$$\left(1 - \frac{r^2}{6\tau_{min}^2}\right)^d \mathcal{H}^d(A) \leq \mathcal{H}^d(\exp_q(A)) \leq \left(1 + \frac{r^2}{\tau_{min}^2}\right)^d \mathcal{H}^d(A).$$

Proposition III.22 collects results that were already known and scattered in the literature. (i) is stated in Proposition 2.1 in [NSW08], yielding (ii) from Corollary 1.4 in [AB06a]. (iii) follows using (i) again and the Gauss equation [dC92, p. 130]. (iv) is derived from (iii) by a direct application of Lemma 8 in [DVW15]. (v) follows from (iv) and Lemma 6 in [ACLZ17].

From the above results, we see that submanifolds with reach bounded away from zero by a fixed constant behave well — quantitatively — with respect to standard differential geometry quantities. It will then provide good statistical models.

III.3.3 Sampling on Submanifolds with Reach Bounded Away from Zero

Let us come back to consider M as generating data. Here, if $\tau_M \geq \tau_{min} > 0$ and that we sample roughly uniformly on M, Proposition III.22 yields that, at scale τ_{min}, the distribution of points roughly behaves like the d-dimensional Lebesgue measure.

Lemma III.23. *Let $M \subset \mathbb{R}^D$ be a compact d-dimensional submanifold with reach $\tau_M \geq \tau_{min} > 0$. Let P be a probability distribution that has a density $f_{min} \leq f \leq f_{max}$ with respect to the volume measure on M. Then for all $r \leq \tau_{min}/4$ and x in M,*

$$c_d f_{min} r^d \leq P(\mathcal{B}(x, r)) \leq C_d f_{max} r^d,$$

for some $c_d, C_d > 0$. As a consequence, if $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ is an i.i.d. n-sample of P, then for for $r \leq \tau_{min}/2$,

$$\mathbb{P}\left(d_H(M, \mathbb{X}_n) > r\right) \leq \frac{4^d}{c_d f_{min} r^d} \exp\left(-n c_d f_{min} r^d\right).$$

In particular, letting $h = \left(\frac{C'_d k}{f_{min}} \frac{\log n}{n}\right)^{1/d}$ with C'_d large enough, the following holds. For n large enough so that $h \leq \tau_{min}/2$, with probability at least $1 - \left(\frac{1}{n}\right)^{k/d}$,

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

The proof of Lemma III.23 is a straightforward combination of Lemma III.21 and Proposition III.22 (v), yielding standardness of the distribution P at scale $\tau_{min}/4$ (Definition B.3). Conclude with Proposition III.14 for the Hausdorff distance bounds. It is worth noting that here, the assumption $f \leq f_{max}$ is not necessary to derive deviations on $d_H(M, \mathbb{X}_n)$.

Lemma III.23 is a key result for all the inference geometry results we will derive in this thesis. Indeed, if one wants to estimate geometric or topological quantities associated to M from point a cloud $\mathbb{X}_n$, then the least we can ask is that $\mathbb{X}_n$ covers M densely enough. If not, there is not hope to recover any geometric information from zones that $\mathbb{X}_n$ does not span in M.

III.3.4 Implicit Constraints under Reach Regularity Condition

The models considered in this thesis all satisfy the assumptions of Proposition III.23 (with $f_{max} = \infty$ for Chapter V). By now, it is worth noting that although the assumptions may seem weak, they actually provide extra information about the geometry of the studied submanifolds.

First, from Proposition III.23 and Proposition III.13 we see that the diameter of submanifolds cannot be arbitrarily large.

Lemma III.24. *Let $M \subset \mathbb{R}^D$ be a compact connected d-dimensional submanifold. Let P be a probability distribution having support M with a density $f \geq f_{min}$ with respect to the Hausdorff measure on M. Then,*

$$diam(M) \leq \frac{C_d}{\tau_M^{d-1} f_{min}},$$

for some constant $C_d > 0$ depending only on d.

In particular, if $\tau_M \geq \tau_{min} > 0$, then $diam(M) \leq C_d/(\tau_{min}^{d-1} f_{min})$. Hence, if τ_{min} and f_{min} are fixed, one gets a bound on $diam(M)$, although not required explicitly. Furthermore, considering the uniform probability distribution on M — corresponding to $f = \mathcal{H}^d(M)^{-1} = Vol(M)^{-1}$ — notice that we obtain the bound

$$diam(M) \leq C_d \frac{Vol(M)}{\tau_M^{d-1}},$$

which is interesting for itself.

Let us carry on with a bound relating directly reach and diameter. One can see the following Proposition III.25 as a complementary constraint to Lemma III.24, as it states that the reach is a lower bound on the diameter, up to universal constants.

Proposition III.25. *If $K \subset \mathbb{R}^D$ is not homotopy equivalent to a point,*

$$\tau_K \leq \sqrt{\frac{D}{2(D+1)}} diam(K).$$

The proof of Proposition III.25 is a straightforward combination of Lemma A.1 and Lemma A.2, that we defer to the appendix. Notice that the assumption that K is not homotopy equivalent to a point cannot be released. Indeed, if one takes K to be a spherical cap of radius 1 and height $0 < h < 1$, then $\tau_K = 1$ although $diam(K)$ goes to 0 as h goes to 0.

Finally, notice that if M is a submanifold of dimension d, then Theorem 3.26 in [Hat02] asserts that it has a non trivial homology group of dimension d over $\mathbb{Z}/2\mathbb{Z}$, so that it cannot be homotopy equivalent to a point. Therefore, combining Lemma III.24 and Proposition III.25 yields the following.

Proposition III.26. *Let $M \subset \mathbb{R}^D$ be a compact connected d-dimensional submanifold. Let P be a probability distribution having support M with a density $f \geq f_{min}$ with respect to the Hausdorff measure on M. Then,*

$$\tau_M^d \leq \frac{C_d}{f_{min}},$$

for some constant $C_d > 0$ depending only on d.

As a consequence, if one considers a statistical model composed of distributions P with support being submanifolds M with reach $\tau_M \geq \tau_{min} > 0$, and with densities $f \geq f_{min} > 0$, then both diameter $diam(M)$ and reach τ_M are bounded away from 0 and ∞ by constants depending only on d, τ_{min} and f_{min}.

III.4 Angles Between Vector Subspaces

In this thesis, we consider the question of estimating tangent spaces of a submanifold. For that, we need a notion of angle between vector subspaces of dimension greater than one. In addition, angles between subspaces will be a useful technical tool in several derivations. As it is not standard in the literature, let us define the notion of angles we use.

For two vector subspaces U, V of $\mathbb{R}^D$, we will measure the angle between them by

$$\angle(U, V) = \|\pi_V - \pi_U\|_{op}. \tag{III.27}$$

In other words, we identify a subspace to its orthogonal projector, yielding a metric induced by matrix norms. Any other norm on the space of matrix would give a good notion of angle. However, some of these may depend cruedly on the ambient dimension D, which is not the case for the operator norm, since

$$\|\pi_V - \pi_U\|_{op} = \sup_{x \in U+V} \frac{\|\pi_V(x) - \pi_U(x)\|}{\|x\|},$$

where $\dim(U+V) \leq \dim(U)+\dim(V)$. Note that the Frobenius norm $\|A\|_{\mathcal{F}} = \sqrt{\operatorname{trace}(A^t A)}$ yields a notion of angle equivalent to $\angle(U, V)$ up to constants independent of D, since

$$\begin{aligned}\|\pi_V - \pi_U\|_{op} \leq \|\pi_V - \pi_U\|_{\mathcal{F}} &\leq \sqrt{\operatorname{rank}(\pi_V - \pi_U)}\, \|\pi_V - \pi_U\|_{op} \\ &\leq \sqrt{\dim(U+V)}\, \|\pi_V - \pi_U\|_{op}.\end{aligned}$$

Another popular definition of angle is the principal angle [BGO09]. The principal angles between U and V is

$$\sin\theta(U, V) = \max_{\substack{u \in U \\ \|u\|=1}} \min_{\substack{v \in V \\ \|v\|=1}} \sqrt{1 - < u, v >^2} \tag{III.28}$$

Note that $\theta(U, V)$ has no reason to be equal to $\theta(V, U)$ in general. For instance, if $U \subset V$ and $\dim(U) < \dim(V)$, we have $\theta(U, V) = 0$ while $\theta(V, U) = \pi/2$. These two notions of angle are closely related, as stated in the following proposition, which proof follows from Theorem 2.6.1 in [GVL96] and Theorem 3.6 in [KA02].

Proposition III.29. *For all vector subspaces U, V of $\mathbb{R}^D$, $\sin\theta(U, V) = \|\pi_{V^\perp} \circ \pi_U\|_{op}$. The principal angle satisfies $\theta(U^\perp, V^\perp) = \theta(V, U)$, and if $\dim(U) = \dim(V)$, then $\theta(U, V) = \theta(V, U)$. Moreover, in his case,*

$$\angle(U, V) = \sin\theta(U, V).$$

Furthermore, there exists a rotation $R_{U \to V}$ of $\mathbb{R}^D$ such that $R_{U \to V} : U \to V$ is bijective, and $\|R_{U \to V} - I_D\|_{op} \leq 2\sin(\theta(U, V)/2)$.

To sum up, $0 \leq \angle(U, V) \leq 1$, and if U and V have the same dimension, their angle $\angle(U, V)$ coincides with the sine $\sin\theta(U, V)$ of the principal angle. It is zero if and only if $U = V$, and one if and only if $U \cap V^\perp \neq \{0\}$. Finally, let us mention that the angle $\angle(U, V) = \sin\theta(U, V)$ when $\dim(U) = \dim(V) = d$ can be computed in $\mathcal{O}(Dd^2)$ using a singular value decomposition [GVL96, §12.4.3].

Let us illustrate these notions of angles with the variations of tangent spaces of a submanifold with reach bounded away from zero. We saw in Section III.3.2 that τ_M^{-1} yields a bound on the curvature of M. Such a bound casts in terms of tangent space variations as follows.

Proposition III.30. *Let $M \subset \mathbb{R}^D$ be a submanifold with $\tau_M \geq \tau_{min} > 0$. Then, for all $p, q \in M$,*

$$\sin\theta\left(T_pM, < q - p >\right) \leq \frac{\|q-p\|}{2\tau_{min}},$$

where $< u >$ denotes the span of $u \in \mathbb{R}^D$. Furthermore, for $\|q-p\| \leq \tau_{min}/2$,

$$\angle(T_pM, T_qM) \leq 2\frac{\|q-p\|}{\tau_{min}}\sqrt{1 - \left(\frac{\|q-p\|}{\tau_{min}}\right)^2}.$$

The first statement is (III.19) rephrased in terms of angle, and the second statement can be found in Lemma 3.4 in [BSW09].

Appendix A

Proofs for Chapter III

Content

A.1 Hausdorff Distance

Proof for Definition III.2. It is clear from the definition that $d_H(\cdot,\cdot)$ is finite on compact sets, and symmetric. Moreover, if $d_H(A,B)=0$, then for all $a\in A$, $d(a,B)=0$. Since B is a closed subset, we get $A\subset B$. Symmetrically, we get $B\subset A$, which shows that d_H is separated. Let now $A,B,C\in\mathcal{K}(\mathcal{D})$. Since $d(\cdot,B)$ is 1-Lipschitz, for all $a\in A$ and $c\in C$, $d(a,B)\leq d(c,B)+d(a,c)$. By definition, $d(c,B)\leq d_H(C,B)$. Hence,

$$\begin{aligned} d(a,B) &\leq d_H(C,B)+\inf_{c\in C} d(a,c)\\ &= d_H(C,B)+d(a,C)\\ &\leq d_H(C,B)+d_H(A,C), \end{aligned}$$

so that $\sup_{a\in A} d(a,B)\leq d_H(C,B)+d_H(A,C)$. Symmetrically, we get $\sup_{b\in B} d(b,A)\leq d_H(C,A)+d_H(B,C)$, which gives the triangle inequality $d_H(A,B)\leq d_H(A,C)+d_H(C,B)$. □

Proof of Proposition III.4. By definition, for all $a\in A$, $d(a,B)\leq d_H(A,B)$, which yields $B^{d_H(A,B)}\supset A$. Symmetrically, $A^{d_H(A,B)}\supset B$, and hence

$$d_H(A,B)\geq \inf\left\{r>0, A^r\supset B \text{ and } B^r\supset A\right\}.$$

Conversely, without loss of generality, there exists a point $a_0\in A$ such that $d(a_0,B)=d_H(A,B)$. Hence, for all $r<d_H(A,B)$, $a_0\notin B^r$ and in particular $B^r\not\supseteq A$. Hence the result. □

Proof of Proposition III.5. For all $x\in A$, $d(x,A)=0$, so that

$$\sup_{a\in A} d(a,B)=\sup_{x\in A} d(x,B)-d(x,A)\leq \sup_{x\in\mathcal{D}} d(x,B)-d(x,A).$$

We now prove the reverse inequality. For $x \in \mathcal{D}$, write $\pi_A(x)$ for any element of A such that $d(x, A) = d(x, \pi_A(x))$. Then, since $d(\cdot, B)$ is 1-Lipschitz,

$$\begin{aligned} d(x, B) - d(x, A) &= d(x, B) - d(x, \pi_A(x)) \\ &\leq d(\pi_A(x), B) \\ &\leq \sup_{a \in A} d(a, B), \end{aligned}$$

which yields the desired reverse bound, and hence

$$\sup_{a \in A} d(a, B) = \sup_{x \in \mathcal{D}} d(x, B) - d(x, A).$$

By symmetry,

$$\sup_{b \in B} d(b, A) = \sup_{x \in \mathcal{D}} d(x, A) - d(x, B).$$

Conclude writing

$$\begin{aligned} \sup_{x \in \mathcal{D}} |d(x, A) - d(x, B)| &= \max \left\{ \sup_{x \in \mathcal{D}} d(x, B) - d(x, A), \sup_{x \in \mathcal{D}} d(x, A) - d(x, B) \right\} \\ &= \max \left\{ \sup_{a \in A} d(a, B), \sup_{b \in B} d(b, A) \right\} \\ &= d_H(A, B). \end{aligned}$$

Finally, if $(\mathcal{D}, d)$ is complete, the closedness of $\mathcal{K}(\mathcal{D})$ in the space of continuous functions is proved in Lemma 3.1.1 of [Bee93]. □

Proof of Proposition III.6. (i) For the direct sense, notice that a dense sequence $\{x_i\}_{i \in \mathbb{N}}$ of $\mathcal{D}$ provides the countable family $\{\cup_{i \in I} \{x_i\}\}_{\text{finite } I \subset \mathbb{N}}$ which is dense in $\mathcal{K}(\mathcal{D})$. Conversely, if $\mathcal{K}(\mathcal{D})$ is separable, so is $\mathcal{D}$, as a closed subset of the metric space $\mathcal{K}(\mathcal{D})$.

(ii) If $(\mathcal{D}, d)$ is compact, then $\mathcal{K}(\mathcal{D}) \simeq \{d(\cdot, A)\}_{A \in \mathcal{K}(\mathcal{D})}$ is an equicontinuous and relatively compact family of $\mathcal{C}(\mathcal{D}, \mathbb{R}_+)$, with $\mathcal{D}$ compact. Hence, it is compact. Conversely, if $\mathcal{K}(\mathcal{D})$ is compact, so is $\mathcal{D}$, as a closed subset of $\mathcal{K}(\mathcal{D})$.

(iii) Follows from the same argument as (ii) by localizing.

(iv) From Proposition III.5, $\mathcal{K}(\mathcal{D})$ is a closed subset of the complete space space $\mathcal{C}(\mathcal{D}, \mathbb{R}_+)$, so it is complete. Conversely, if $\mathcal{K}(\mathcal{D})$ is complete, so is $\mathcal{D}$, as a closed subset of $\mathcal{K}(\mathcal{D})$.

(v) Is a rephrasing of (i) with (iv).

□

Proof of III.7. We actually prove continuity of the considered map, which is stronger than measurability.

(i) It is an isometry.

(ii) To prove that the map $(f, A) \mapsto f(A)$ is jointly measurable, it is sufficient to prove continuous in each variable separately, from Lemma 4.51 in [AB06b]. Fix $A \in \mathcal{K}(\mathbb{R}^D)$. Then for all f, g continous, $d_H(f(A), g(A)) \leq \sup_{x \in A} |f(x) - g(x)|$, which goes to zero when g converges to f on the compact A. Let now f be fixed. Then for all $A \in \mathcal{K}(\mathbb{R}^D)$, consider $K = A^1$ the offset of radius 1 of A. Then f is uniformly continuous on the

compact set K. Hence, for all $\varepsilon > 0$, there exists $\eta > 0$ such that for all $x, y \in K$ such that $\|y - x\| \leq \eta$, $\|f(y) - f(x)\| \leq \varepsilon$. Hence, for $d_H(A, B) \leq \eta \wedge 1$, we get $d_H(f(A), f(B)) \leq \varepsilon$, which proves continuity of $B \mapsto f(B)$ at A, and concludes the proof.

(iii) Writting $r = \max\{d_H(A_1, A_2), d_H(B_1, B_2)\}$, we have $(A_1 \cup B_1)^r = A_1^r \cup B_1^r \supset A_2 \cup B_2$. Symmetrically, $(A_2 \cup B_2)^r \supset A_1 \cup A_2$, so that

$$d_H(A_1 \cup B_1, A_2 \cup B_2) \leq \max\{d_H(A_1, A_2), d_H(B_1, B_2)\}.$$

(iv) For any convex combination $\bar{a} = \sum_i \lambda_i a_i \in conv(A)$ of elements of A, considering convex combinations $\bar{b} = \sum_i \lambda_i b_i$ for $b_i \in B$ clearly yields

$$d(\bar{a}, conv(B)) \leq \sum_i \lambda_i d(a_i, B) \leq \sum_i \lambda_i d_H(A, B) = d_H(A, B).$$

Symmetrically, we obtain $d(\bar{b}, conv(A))$ for all $\bar{b} \in conv(B)$. Hence,

$$d_H(conv(A), conv(B)) \leq d_H(A, B).$$

□

A.2 Standard Measures

Proof of Proposition III.11. The proof follows that of Lemma 10 in [CGLM15]. Let $\mathbf{y} = \{y_1, \ldots, y_k\}$ ($k = \mathrm{pk}\,(Supp(\mu), r)$)be a maximal r-packing of $Supp(\mu)$. We have

$$\begin{aligned} 1 = \mu\,(Supp(\mu)) &\geq \mu\,(\cup_{i=1}\mathcal{B}(y_i, r)) \\ &\geq \sum_{i=1}^{k} \mu\,(\mathcal{B}(y_i, r)) \\ &\geq kar^b = \mathrm{pk}\,(Supp(\mu), r)\, ar^b, \end{aligned}$$

hence the first result. The bound on $\mathrm{cv}\,(Supp(\mu), r)$ then follows from (III.10). □

Proof of Lemma III.12. Let $p, q \in K$ and $\gamma : [0, 1] \to K$ be a continuous path joining $\gamma(0) = p$ and $\gamma(1) = q$. Writing $N = \mathrm{cv}(K, \varepsilon)$, let $x_1, \ldots, x_N \in \mathbb{R}^D$ be the centers of a covering of K by open balls of radii ε. We let U_i denote $\{t, \|\gamma(t) - x_i\| < \varepsilon\} \subset [0, 1]$. By construction of the covering, there exists $x_{(1)} \in \{x_1, \ldots, x_N\}$ such that $\left\|p - x_{(1)}\right\| < \varepsilon$. Then $U_{(1)} \ni \gamma(0) = p$ is a non-empty open subset of $[0, 1]$, so that $t_{(1)} = \sup U_{(1)}$ is positive. If $t_{(1)} = 1$, then $\left\|q - x_{(1)}\right\| \leq \varepsilon$, and in particular $\|q - p\| \leq 2\varepsilon$. If $t_{(1)} < 1$, since $U_{(1)}$ is an open subset of $[0, 1]$, we see that $\gamma(t_{(1)}) \notin U_{(1)}$. But $\cup_{i=1}^{N} U_i$ is an open cover of $[0, 1]$, which yields the existence $U_{(2)}$ such that $\gamma(t_{(1)}) \in U_{(2)}$, and for all $t < t_{(1)}$, $\gamma(t) \notin U_{(2)}$. Then consider $t_{(2)} = \sup U_{(2)}$, and so on. Doing so, we build by induction a sequence of numbers $0 < t_{(1)} < \ldots < t_{(k)} \leq 1$ and distinct centers $x_{(1)}, \ldots, x_{(k)} \in \{x_1, \ldots, x_N\}$ ($k \leq N$) such that $\left\|p - x_{(1)}\right\| < \varepsilon$, $\left\|q - x_{(k)}\right\| \leq \varepsilon$, with $\left\|\gamma(t_{(i)}) - x_{(i)}\right\| \leq \varepsilon$ for $1 \leq i \leq k$ and $\left\|\gamma(t_{(i)}) - x_{(i+1)}\right\| < \varepsilon$ for $1 \leq i \leq k - 1$. In particular, $\left\|x_{(i)} - x_{(i+1)}\right\| \leq 2\varepsilon$ for all $1 \leq i \leq k - 1$. To conclude, write

$$\begin{aligned} \|p - q\| &\leq \left\|p - x_{(1)}\right\| + \left\|x_{(1)} - x_{(k)}\right\| + \left\|q - x_{(k)}\right\| \\ &\leq \varepsilon + \sum_{i=1}^{k-1} \left\|x_{(i)} - x_{(i+1)}\right\| + \varepsilon \\ &\leq 2k\varepsilon \leq 2\varepsilon \mathrm{cv}(K, \varepsilon). \end{aligned}$$

Since this bound holds for all $p, q \in K$, we get the announced bound on the diameter of K. □

Proof of Proposition III.14. Since $\mathbb{X}_n \subset Supp(\mu)$ with probability one, the Hausdorff distance between $\mathbb{X}_n$ and $Supp(\mu)$ writes

$$d_H\big(Supp(\mu), \mathbb{X}_n\big) = \sup_{x \in Supp(\mu)} \min_{1 \le j \le n} d(x, X_j).$$

For some $\delta > 0$ to be chosen later, consider a minimal δ-covering $\mathbf{x} = \{x_1, \dots, x_k\}$ of $Supp(\mu)$ ($k = \mathrm{cv}(Supp(\mu), \delta)$). For all $x \in Supp(\mu)$, there exists some $x_i \in \mathbf{x}$ such that $d(x, x_i) \le \delta$. Hence,

$$\min_{1 \le j \le n} d(x, X_j) \le \delta + \max_{1 \le i \le k} \min_{1 \le j \le n} d(x_i, X_j).$$

As a consequence,

$$\begin{aligned}
\mathbb{P}\left(d_H\big(Supp(\mu), \mathbb{X}_n\big) > r\right) &\le \mathbb{P}\left(\max_{1 \le i \le k} \min_{1 \le j \le n} d(x_i, X_j) > r - \delta\right) \\
&\le \sum_{i=1}^{k} \mathbb{P}\left(\min_{1 \le j \le n} d(x_i, X_j) > r - \delta\right)
\end{aligned}$$

But whenever $r - \delta \le r_0$, for all $1 \le i \le k$,

$$\begin{aligned}
\mathbb{P}\left(\min_{1 \le j \le n} d(x_i, X_j) > r - \delta\right) &= \prod_{1 \le j \le n} \mathbb{P}\left(d(x_i, X_j) > r - \delta\right) \\
&= \left(1 - \mu\left(\mathcal{B}(x_i, r - \delta)\right)\right)^n \\
&\le \Big(1 - a(r - \delta)^b\Big)^n \le \exp\Big(-na(r - \delta)^b\Big).
\end{aligned}$$

Therefore, Proposition III.11 yields for all $\delta \le 2r_0$ such that $r - \delta \le r_0$,

$$\begin{aligned}
\mathbb{P}\left(d_H\big(Supp(\mu), \mathbb{X}_n\big) > r\right) &\le \mathrm{cv}(Supp(\mu), \delta) \exp\Big(-na(r - \delta)^b\Big) \\
&\le \frac{2^b}{a\delta^b} \exp\Big(-na(r - \delta)^b\Big).
\end{aligned}$$

Setting $\delta = r/2$ yields the announced result. □

A.3 Constraints Given by the Reach

This section includes two intermediate results yielding Proposition III.25. We let $conv(\cdot)$ denote the closed convex hull of a set.

Lemma A.1. *For all $K \subset \mathbb{R}^D$, $d_H\left(K, conv(K)\right) \le \sqrt{\frac{D}{2(D+1)}} diam(K)$.*

Proof of Lemma A.1. It is a straightforward corollary of Jung's Theorem 2.10.41 in [Fed69], which states that K is contained in a (unique) closed ball with (minimal) radius at most $\sqrt{\frac{D}{2(D+1)}} diam(K)$. □

Lemma A.2. *If $K \subset \mathbb{R}^D$ is not homotopy equivalent to a point, $\tau_K \le d_H\left(K, conv(K)\right)$.*

Proof of Lemma A.2. Let us prove the contrapositive. Assume that $\tau_K > d_H(K, conv(K))$. Then,

$$conv(K) \subset \bigcup_{x \in K} \mathcal{B}(x, d_H(K, conv(K))) \subset \bigcup_{x \in K} \mathring{\mathcal{B}}(x, \tau_K) \subset Med(K)^c.$$

Therefore, the map $\pi_K : conv(K) \to K$ is well defined and continuous, so that K is a retract of $conv(K)$ (see Chapter 0 in [Hat02]). Therefore, K is homotopy equivalent to a point, since the convex set $conv(K)$ is. □

Chapter IV

Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction

Abstract

We consider the problem of optimality in manifold reconstruction. A random sample $\mathbb{X}_n = \{X_1, \dots, X_n\} \subset \mathbb{R}^D$ composed of points close to a d-dimensional submanifold M, with or without outliers drawn in the ambient space, is observed. Based on the Tangential Delaunay Complex [BG14], we construct an estimator $\hat{M}$ that is ambient isotopic and Hausdorff-close to M with high probability. The estimator $\hat{M}$ is built from existing algorithms. In a model with additive noise of small amplitude, we show that this estimator is asymptotically minimax optimal for the Hausdorff distance over a class of submanifolds satisfying a reach constraint. Therefore, even with no *a priori* information on the tangent spaces of M, our estimator based on Tangential Delaunay Complexes is optimal. This shows that the optimal rate of convergence can be achieved through existing algorithms. A similar result is also derived in a model with outliers. A geometric interpolation result is derived, showing that the Tangential Delaunay Complex is stable with respect to noise and perturbations of the tangent spaces. In the process, a decluttering procedure and a tangent space estimator both based on local principal component analysis (PCA) are studied.

Content

IV.1 Introduction

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

The computational geometry community has now been active on manifold reconstruction for many years, mainly in deterministic frameworks. In dimension 3, [Dey07] provides a survey of the state of the art. In higher dimension, the employed methods rely on variants of the ambient Delaunay triangulation [CDR05, BG14]. The geometric and topological guarantees are derived under the assumption that the point cloud — fixed and nonrandom — densely samples M at scale ε, with ε small enough or going to 0.

In the statistics literature, most of the attention has been paid to approximation guarantees, rather than topological ones. The approximation bounds are given in terms of the sample size n, that is assumed to be large enough or going to infinity. To derive these bounds, a broad variety of assumptions on M have been considered. For instance, if M is a bounded convex set and $\mathbb{X}_n$ does not contain outliers, a natural idea is to consider the convex hull $\hat{M} = \mathrm{conv}(\mathbb{X}_n)$ to be the estimator. $\mathrm{conv}(\mathbb{X}_n)$ provides optimal rates of approximation for several loss functions [MT95, DW96]. These rates depend crudely on the regularity of the boundary of the convex set M. In addition, $\mathrm{conv}(\mathbb{X}_n)$ is clearly ambient isotopic to M so that it has both good geometric and topological properties. Generalisations of the notion of convexity based on rolling ball-type assumptions such as r-convexity and reach bounds [CRC04, GPPVW12a] yield rich classes of sets with good geometric properties. In particular, the reach, as introduced by Federer [Fed59], appears to be a key regularity and scale parameter [CCSL06, GPPVW12a, MMS16].

This chapter mainly follows up the two articles [BG14, GPPVW12a], both dealing with the case of a d-dimensional submanifold $M \subset \mathbb{R}^D$ with a reach regularity condition and where the dimension d is known.

On one hand, [BG14] focuses on a deterministic analysis and proposes a provably faithful reconstruction. The authors introduce a weighted Delaunay triangulation restricted to tangent spaces, the so-called Tangential Delaunay Complex. This chapter gives a reconstruction up to ambient isotopy with approximation bounds for the Hausdorff distance along with computational complexity bounds. This work provides a simplicial complex based on the input point cloud and tangent spaces. However, it lacks stability up to now, in the sense that the assumptions used in the proofs of [BG14] do not resist ambient perturbations. Indeed, it heavily relies on the knowledge of the tangent spaces at each point and on the absence of noise.

On the other hand, [GPPVW12a] takes a statistical approach in a model possibly corrupted by additive noise, or containing outlier points. The authors derive an estimator that is proved to be minimax optimal for the Hausdorff distance d_H. Roughly speaking, minimax optimality of the proposed estimator means that it performs best in the worst possible case up to numerical constants, when the sample size n is large enough. Although theoretically optimal, the proposed estimator appears to be intractable in practice. At last, [MMS16] proposes a manifold estimator based on local linear patches that is tractable but fails to achieve the optimal rates.

Contribution

Our main contributions (Theorems IV.7, IV.8 and IV.9) make a two-way link between the approaches of [BG14] and [GPPVW12a].

From a geometric perspective, Theorem IV.7 shows that the Tangential Delaunay Complex of [BG14] can be combined with local PCA to provide a manifold estimator that is optimal in the sense of [GPPVW12a]. This remains possible even if data is corrupted with additive noise of small amplitude. Also, Theorems IV.8 and IV.9 show that, if outlier points are present (clutter noise), the Tangential Delaunay Complex of [BG14] still yields the optimal rates of [GPPVW12a], at the price of an additional decluttering procedure.

From a statistical point of view, our results show that the optimal rates described in [GPPVW12a] can be achieved by a tractable estimator $\hat{M}$ that (1) is a simplicial complex of which vertices are the data points, and (2) such that $\hat{M}$ is ambient isotopic to M with high probability.

In the process, a stability result for the Tangential Delaunay Complex (Theorem IV.14) is proved. Let us point out that this stability is derived using an interpolation result (Theorem IV.11) which is interesting in its own right. Theorem IV.11 states that if a point cloud $\mathcal{X}$ lies close to a submanifold M, and that estimated tangent spaces at each sample point are given, then there is a submanifold M' (ambient isotopic, and close to M for the Hausdorff distance) that interpolates $\mathcal{X}$, with T_pM' agreeing with the estimated tangent spaces at each point $p \in \mathcal{X}$. Moreover, the construction can be done so that the reach of M' is bounded in terms of the reach of M, provided that $\mathcal{X}$ is sparse, points of $\mathcal{X}$ lie close to M, and error on the estimated tangent spaces is small. Hence, Theorem IV.11 essentially allows to consider a noisy sample with estimated tangent spaces as an exact sample with exact tangent spaces on a proxy submanifold. This approach can provide stability for any algorithm that takes point cloud and tangent spaces as input, such as the so-called *cocone* complex [CDR05].

Outline

This chapter deals with the case where a sample $\mathbb{X}_n = \{X_1, \dots, X_n\} \subset \mathbb{R}^D$ of size n is randomly drawn on/around M. First, the statistical framework is described (Section IV.2.1) together with minimax optimality (Section IV.2.2). Then, the main results are stated (Section IV.2.3).

Two models are studied, one where $\mathbb{X}_n$ is corrupted with additive noise, and one where $\mathbb{X}_n$ contains outliers. We build a simplicial complex $\hat{M}_{\texttt{TDC}}(\mathbb{X}_n)$ ambient isotopic to M and we derive the rate of approximation for the Hausdorff distance $d_H(M, \hat{M}_{\texttt{TDC}})$, with bounds holding uniformly over a class of submanifolds satisfying a reach regularity condition. The derived rate of convergence is minimax optimal if the amplitude σ of the additive noise is small. With outliers, similar estimators $\hat{M}_{\texttt{TDC}\delta}$ and $\hat{M}_{\texttt{TDC}+}$ are built. $\hat{M}_{\texttt{TDC}}$, $\hat{M}_{\texttt{TDC}\delta}$ and $\hat{M}_{\texttt{TDC}+}$ are based on the Tangential Delaunay Complex (Section IV.3), that is first proved to be stable (Section IV.4) via an interpolation result. A method to estimate tangent spaces and to remove outliers based on local Principal Component Analysis (PCA) is proposed (Section IV.5). We conclude with general remarks and possible extensions (Section IV.6). For ease of exposition, all the proofs are placed in the appendix.

Notation

In what follows, we consider a compact d-dimensional submanifold without boundary $M \subset \mathbb{R}^D$ to be reconstructed. For all $p \in M$, T_pM designates the tangent space of M at

p. Tangent spaces will either be considered vectorial or affine depending on the context. The standard inner product in $\mathbb{R}^D$ is denoted by $\langle \cdot, \cdot \rangle$ and the Euclidean distance $\|\cdot\|$. We let $\mathcal{B}(p, r)$ denote the closed Euclidean ball of radius $r > 0$ centered at p. We let $\wedge$ and $\vee$ denote respectively the minimum and the maximum of real numbers. As introduced in [Fed59], the reach of M, denoted by τ_M is the maximal offset radius for which the projection π_M onto M is well defined. Denoting by $d(\cdot, M)$ the distance to M, the *medial axis* of M $\text{med}(M) = \{x \in \mathbb{R}^D | \exists a \neq b \in M, \|x - a\| = \|x - b\| = d(x, M)\}$ is the set of points which have at least two nearest neighbors on M. Then, $\tau_M = \inf_{p \in M} d(p, \text{med}(M))$. We simply write π for π_M when there is no possibility of confusion. For any smooth function $\Phi : \mathbb{R}^D \to \mathbb{R}^D$, we let $d_a\Phi$ and $d_a^2\Phi$ denote the first and second order differentials of Φ at $a \in \mathbb{R}^D$. For a linear map A, A^t designates its transpose. Let $\|A\|_{\text{op}} = \sup_x \frac{\|Ax\|}{\|x\|}$ and $\|A\|_{\mathcal{F}} = \sqrt{\text{trace}\,(A^t A)}$ denote respectively the operator norm induced by the Euclidean norm and the Frobenius norm. The distance between two linear subspaces $U, V \subset \mathbb{R}^D$ of the same dimension is measured by the sine

$$\angle(U, V) = \max_{u \in U} \max_{v' \in V^\perp} \frac{\langle u, v' \rangle}{\|u\| \, \|v'\|} = \|\pi_U - \pi_V\|_{\text{op}}$$

of their largest principal angle, as defined Section III.4. The Hausdorff distance between two compact subsets K, K' of $\mathbb{R}^D$ is denoted by $d_H(K, K') = \sup_{x \in \mathbb{R}^D} |d(x, K) - d(x, K')|$. Finally, we let $\cong$ denote the ambient isotopy relation in $\mathbb{R}^D$.
Throughout, C_α will denote a generic constant depending on the parameter α. For clarity's sake, c_α and K_α may also be used when several constants are involved.

IV.2 Minimax Risk and Main Results

IV.2.1 Statistical Model

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

Let $\mathcal{C}^2_{\tau_{min}}$ be the set of all compact d-dimensional connected submanifolds $M \subset \mathbb{R}^D$ without boundary satisfying $\tau_M \geq \tau_{min}$. The reach assumption is crucial to avoid arbitrarily curved and pinched shapes [CRC04]. From a reconstruction point of view, τ_{min} gives a minimal feature size on M, and then a minimal scale for geometric information. Every $M \in \mathcal{C}^2_{\tau_{min}}$ inherits a measure induced by the d-dimensional Hausdorff measure on $\mathbb{R}^D \supset M$. We denote this induced measure by v_M. Beyond the geometric restrictions induced by the lower bound τ_{min} on the reach, it also requires the natural measure v_M to behave like a d-dimensional measure, up to uniform constants. Namely, v_M satisfies the (a, d)-standard property of Chapter III, at scale smaller than τ_{min}, and with $a = a_d$. Denote by $\mathcal{U}_M(f_{min}, f_{max})$ the set of probability distributions Q having a density f with respect to v_M such that $0 < f_{min} \leq f(x) \leq f_{max} < \infty$ for all $x \in M$. In particular, notice that such distributions $Q \in \mathcal{U}_M(f_{min}, f_{max})$ all have support M. Roughly speaking, when $Q \in \mathcal{U}_M(f_{min}, f_{max})$, points are drawn almost uniformly on M. This is to ensure that the sample visits all the areas of M with high probability. The noise-free model $\mathcal{P}^2_{\tau_{min}}(f_{min}, f_{max})$ consists of the set of all these almost uniform measures on submanifolds of dimension d having reach greater than a fixed value $\tau_{min} > 0$.

Definition IV.1 (Noise-free model)**.** *In what follows, we write*

$$\mathcal{P}^2_{\tau_{min}}(f_{min}, f_{max}) = \bigcup_{M \in \mathcal{C}^2_{\tau_{min}}} \mathcal{U}_M(f_{min}, f_{max}).$$

We do not explicitly impose a bound on the diameter of M. Actually, a bound is implicitly present in the model according to Lemma III.24. Let us state the bound here for sake of completeness.

Lemma IV.2. *There exists $C_d > 0$ such that for all $Q \in \mathcal{P}^2_{\tau_{min}}(f_{min}, f_{max})$ with associated M,*

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

Random variables with distribution belonging to the noise-free model $\mathcal{P}^2_{\tau_{min}}(f_{min}, f_{max})$ lie exactly on the submanifold of interest M. A more realistic model should allow some measurement error, as illustrated by Figure IV.1a. We formalize this idea with the following additive noise model.

Definition IV.3 (Additive noise model)**.** *For $\sigma < \tau_{min}$, we let $\mathcal{P}^2_{\tau_{min},\sigma}(f_{min}, f_{max})$ denote the set of distributions of random variables $X = Y + Z$, where Y has distribution $Q \in \mathcal{P}^2_{\tau_{min}}(f_{min}, f_{max})$, and $\|Z\| \leq \sigma$ almost surely.*

Let us emphasize that we do not require Y and Z to be independent, nor Z to be orthogonal to $T_Y M$, as done for the "perpendicular" noise model of [NSW08, GPPVW12a]. This model is also slightly more general than the one considered in [MMS16]. Notice that the noise-free model can be thought of as a particular instance of the additive noise model, since $\mathcal{P}^2_{\tau_{min}}(f_{min}, f_{max}) = \mathcal{P}^2_{\tau_{min},\sigma=0}(f_{min}, f_{max})$.

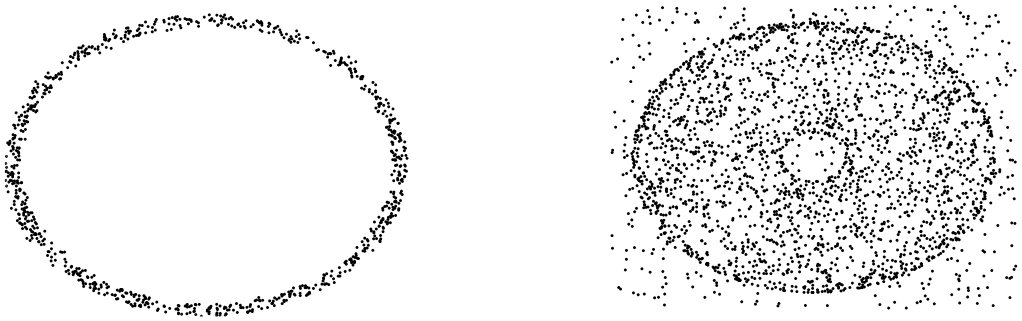
Eventually, we may include distributions contaminated with outliers uniformly drawn in a ball $\mathcal{B}_0$ containing M, as illustrated in Figure IV.1b. Up to translation, we can always assume that $M \ni 0$. To avoid boundary effects, $\mathcal{B}_0$ will be taken to contain M amply, so that the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

Definition IV.4 (Model with outliers/Clutter noise model)**.** *For $0 < f_{min} \leq f_{max} < \infty$, $0 < \beta \leq 1$, and $\tau_{min} > 0$, we define $\mathcal{P}^2_{\tau_{min},\beta}(f_{min}, f_{max})$ to be the set of mixture distributions*

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

where $Q \in \mathcal{P}^2_{\tau_{min}}(f_{min}, f_{max})$ has support M such that $0 \in M$, and $U_{\mathcal{B}_0}$ is the uniform distribution on $\mathcal{B}_0 = \mathcal{B}(0, K_0)$.

Alternatively, a random variable X with distribution $P \in \mathcal{P}^2_{\tau_{min},\beta}(f_{min}, f_{max})$ can be represented as $X = VX' + (1 - V)X''$, where $V \in \{0, 1\}$ is a Bernoulli random variable with parameter β, X' has distribution in $\mathcal{P}^2_{\tau_{min}}(f_{min}, f_{max})$ and X'' has a uniform distribution over $\mathcal{B}_0$, and such that V, X', X'' are independent. In particular for $\beta = 1$, $\mathcal{P}^2_{\tau_{min},\beta=1}(f_{min}, f_{max}) = \mathcal{P}^2_{\tau_{min}}(f_{min}, f_{max})$.



(a) Circle with noise: $d = 1$, $D = 2$, $\sigma > 0$. *(b)* Torus with outliers: $d = 2$, $D = 3$, $\beta < 1$.

Figure IV.1 – Point clouds $\mathbb{X}_n$ drawn from distributions in $\mathcal{P}^2_{\tau_{min},\sigma}(f_{min}, f_{max})$ (left) and $\mathcal{P}^2_{\tau_{min},\beta}(f_{min}, f_{max})$ (right).

IV.2.2 Minimax Risk

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

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

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

Definition IV.5. *An estimator $\hat{M}$ is said to be (asymptotically) minimax optimal over $\mathcal{D}$ if*

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

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

Studying a minimax rate of convergence is twofold. On one hand, deriving an upper bound on R_n boils down to provide an estimator and to study its quality uniformly on $\mathcal{D}$. On the other hand, bounding R_n from below amounts to study the worst possible case in $\mathcal{D}$. This part is usually achieved with standard Bayesian techniques [LC73]. For the models considered in the present paper, the rates were given in [GPPVW12a, KZ15].

Theorem IV.6 (Theorem 3 of [KZ15])**.** *We have,*

$$R_n\left(\mathcal{P}^2_{\tau_{min}}(f_{min}, f_{max})\right) \asymp \left(\frac{\log n}{n}\right)^{2/d}, \qquad \text{(Noise-free)}$$

and for $0 < \beta \leq 1$ *fixed,*

$$R_n\left(\mathcal{P}^2_{\tau_{min},\beta}(f_{min}, f_{max})\right) \asymp \left(\frac{\log n}{\beta n}\right)^{2/d}. \qquad \text{(Clutter noise)}$$

Since the additive noise model $\mathcal{P}^2_{\tau_{min},\sigma}(f_{min}, f_{max})$ has not yet been considered in the literature, the behavior of the associated minimax risk is not known. Beyond this theoretical result, an interesting question is to know whether these minimax rates can be achieved by a tractable algorithm. Indeed, that proposed in [GPPVW12a] especially rely on a minimization problem over the class of submanifolds $\mathcal{C}^2_{\tau_{min}}$, which is computationally costly. In addition, the proposed estimators are themselves submanifolds, which raises storage problems. Moreover, no guarantee is given on the topology of the estimators. Throughout the present paper, we will build estimators that address these issues.

IV.2.3 Main Results

Let us start with the additive noise model $\mathcal{P}^2_{\tau_{min},\sigma}(f_{min}, f_{max})$, that includes in particular the noise-free case $\sigma = 0$. The estimator $\hat{M}_{\texttt{TDC}}$ is based on the Tangential Delaunay Complex (Section IV.3), with a tangent space estimation using a local PCA (Section IV.5).

Theorem IV.7. *$\hat{M}_{\texttt{TDC}} = \hat{M}_{\texttt{TDC}}(\mathbb{X}_n)$ is a simplicial complex with vertices included in $\mathbb{X}_n$ such that the following holds. There exists $\lambda_{d,f_{min},f_{max}} > 0$ such that if $\sigma \leq \lambda\left(\frac{\log n}{n}\right)^{1/d}$ with $\lambda \leq \lambda_{d,f_{min},f_{max}}$, then*

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

Moreover, for n large enough,

$$\sup_{Q\in\mathcal{P}^2_{\tau_{min},\sigma}(f_{min},f_{max})} \mathbb{E}_{Q^n} d_H(M, \hat{M}_{\texttt{TDC}}) \leq C'_{d,f_{min},f_{max},\tau_{min}} \left\{\left(\frac{\log n}{n}\right)^{2/d} \vee \lambda^2\right\}.$$

It is interesting to note that the constants appearing in Theorem IV.7 do not depend on the ambient dimension D. Since $R_n\left(\mathcal{P}^2_{\tau_{min},\sigma}(f_{min}, f_{max})\right) \geq R_n\left(\mathcal{P}^2_{\tau_{min}}(f_{min}, f_{max})\right)$, we obtain immediately from Theorem IV.7 that $\hat{M}_{\texttt{TDC}}$ achieves the minimax optimal rate $(\log n/n)^{2/d}$ over $\mathcal{P}^2_{\tau_{min},\sigma}(f_{min}, f_{max})$ when $\sigma \leq c_{d,f_{min},f_{max}}(\log n/n)^{2/d}$. Note that the estimator of [MMS16] achieves the rate $(\log n/n)^{2/(d+2)}$ when $\sigma \leq c_{d,f_{min},f_{max}}(\log n/n)^{2/(d+2)}$, so does the estimator of [GPPVW12b] for $\sigma < \tau_{min}$ if the noise is centered and perpendicular to the submanifold. As a consequence, $\hat{M}_{\texttt{TDC}}$ outperforms these two existing procedures whenever $\sigma \ll (\log n/n)^{2/(d+2)}$, with the additional feature of exact topology recovery. Still, for $\sigma \gg (\log n/n)^{1/d}$, $\hat{M}_{\texttt{TDC}}$ may perform poorly compared to [GPPVW12b]. This might be due to the fact that the vertices of $\hat{M}_{\texttt{TDC}}$ are sample points themselves, while for higher noise levels, a pre-process of the data based on local averaging could be more relevant.

In the model with outliers $\mathcal{P}^2_{\tau_{min},\beta}(f_{min}, f_{max})$, with the same procedure used to derive Theorem IV.7 and an additional iterative preprocessing of the data based on local PCA to remove outliers (Section IV.5), we design an estimator of M that achieves a rate as close as wanted to the noise-free rate. Namely, for any positive $\delta < 1/(d(d+1))$, we build $\hat{M}_{\texttt{TDC}\delta}$ that satisfies the following similar statement.

Theorem IV.8. *$\hat{M}_{\texttt{TDC}\delta} = \hat{M}_{\texttt{TDC}\delta}(\mathbb{X}_n)$ is a simplicial complex with vertices included in $\mathbb{X}_n$ such that*

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

Moreover, for n large enough,

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

$\hat{M}_{\texttt{TDC}\delta}$ converges at the rate at least $(\log n/n)^{2/d-2\delta}$, which is not the minimax optimal rate according to Theorem IV.6, but that can be set as close as desired to it. To our knowledge, $\hat{M}_{\texttt{TDC}\delta}$ is the first explicit estimator to provably achieve such a rate in the presence of outliers. Again, it is worth noting that the constants involved in Theorem IV.8 do not depend on the ambient dimension D. The construction and computation of $\hat{M}_{\texttt{TDC}\delta}$ is the same as $\hat{M}_{\texttt{TDC}}$, with an extra pre-processing of the point cloud allowing to remove outliers. This decluttering procedure leads to compute, at each sample point, at most $\log(1/\delta)$ local PCA's, instead of a single one for $\hat{M}_{\texttt{TDC}}$.

From a theoretical point of view, there exists a (random) number of iterations of this decluttering process, from which an estimator $\hat{M}_{\texttt{TDC}+}$ can be built to satisfy the following.

Theorem IV.9. *$\hat{M}_{\texttt{TDC}+} = \hat{M}_{\texttt{TDC}+}(\mathbb{X}_n)$ is a simplicial complex of vertices contained in $\mathbb{X}_n$ such that*

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

Moreover, for n large enough,

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

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

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

IV.3 Tangential Delaunay Complex

Let $\mathcal{X}$ be a finite subset of $\mathbb{R}^D$. In this section, we denote the point cloud $\mathcal{X}$ to emphasize the fact that it is considered nonrandom. For $\varepsilon, \delta > 0$, $\mathcal{X}$ is said to be *ε-dense* in M if $\sup_{x\in M} d(x, \mathcal{X}) \leq \varepsilon$, and *$\delta$-sparse* if $d(p, \mathcal{X} \setminus \{p\}) \geq \delta$ for all $p \in \mathcal{X}$. A *(δ, ε)-net* (of M) is a δ-sparse and ε-dense point cloud.

IV.3.1 Restricted Weighted Delaunay Triangulations

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

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

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

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

Note that for a constant weight assignment $\omega(p) \equiv \omega_0$, $\mathrm{Del}^\omega(\mathcal{X})$ is the usual Delaunay triangulation of $\mathcal{X}$. Under genericity assumptions on $\mathcal{X}$ and bounds on ω, $\mathrm{Del}^\omega(\mathcal{X})$ is an embedded triangulation with vertex set $\mathcal{X}$ [BG14]. The reconstruction method proposed in this paper is based on $\mathrm{Del}^\omega(\mathcal{X})$ for some weights ω to be chosen later. As it is a triangulation of the whole convex hull of $\mathcal{X}$ and fails to recover the geometric structure of M, we take restrictions of it in the following manner.

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

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

In particular, we define the *Tangential Delaunay Complex* $\mathrm{Del}^\omega(\mathcal{X}, T)$ by taking $R = T = \{T_pM\}_{p \in \mathcal{X}}$, the family of tangent spaces taken at the points of $\mathcal{X} \subset M$ [BG14]. $\mathrm{Del}^\omega(\mathcal{X}, T)$ is a pruned version of $\mathrm{Del}^\omega(\mathcal{X})$ where only the simplices with directions close to the tangent spaces are kept. Indeed, T_pM being the best linear approximation of M at p, it is very unlikely for a reconstruction of M to have components in directions normal to T_pM (see Figure IV.2). As pointed out in [BG14], computing $\mathrm{Del}^\omega(\mathcal{X}, T)$ only requires to compute Delaunay triangulations in the tangent spaces that have dimension d. This reduces the computational complexity dependency on the ambient dimension $D > d$. The weight assignment ω gives degrees of freedom for the reconstruction. The extra degree of freedom ω permits to stabilize the triangulation and to remove the so-called *inconsistencies*, the points remaining fixed. For further details, see [BGO09, BG14].

IV.3.2 Guarantees

The following result sums up the reconstruction properties of the Tangential Delaunay Complex that we will use. For more details about it, the reader is referred to [BG14].

Theorem IV.10 (Theorem 5.3 in [BG14])**.** *There exists $\varepsilon_d > 0$ such that for all $\varepsilon \leq \varepsilon_d \tau_{min}$ and all $M \in \mathcal{C}^2_{\tau_{min}}$, if $\mathcal{X} \subset M$ is an $(\varepsilon, 2\varepsilon)$-net, there exists a weight assignment $\omega_* = \omega_{*\mathcal{X},T}$ depending on $\mathcal{X}$ and $T = \{T_pM\}_{p \in \mathcal{X}}$ such that*

- $d_H\left(M, \mathrm{Del}^{\omega_*}(\mathcal{X}, T)\right) \leq C_d \varepsilon^2 / \tau_{min}$,
- *M and $\mathrm{Del}^{\omega_*}(\mathcal{X}, T)$ are ambient isotopic.*

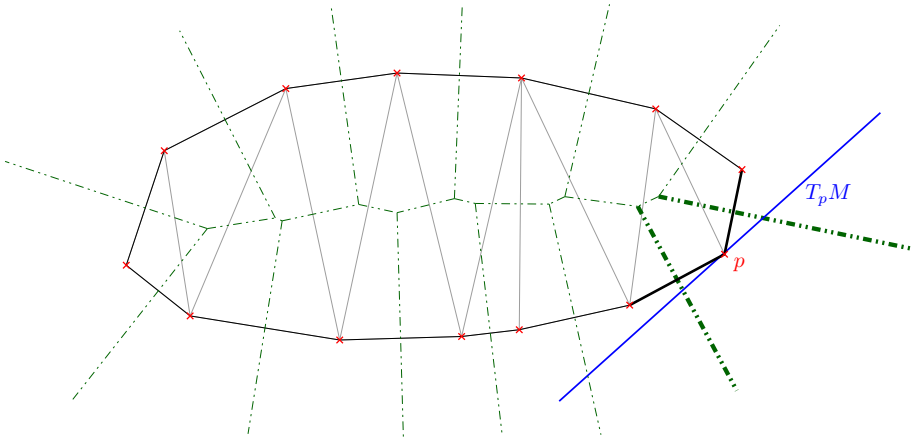


Figure IV.2 – Construction of $\mathrm{Del}^{\omega}(\mathcal{X}, T)$ at p for $\omega \equiv 0$: p has three incident edges in the ambient Delaunay triangulation, but only two (bold) have dual Voronoi face intersecting T_pM.

Computing $\mathrm{Del}^{\omega_*}(\mathcal{X}, T)$ requires to determine the weight function $\omega_* = \omega_{*\mathcal{X},T}$. In [BG14], a greedy algorithm is designed for this purpose and has a time complexity $O\left(Dn^2 + D2^{O(d^2)}n\right)$.

Given an $(\varepsilon, 2\varepsilon)$-net $\mathcal{X}$ for ε small enough, $\mathrm{Del}^{\omega_*}(\mathcal{X}, T)$ recovers M up to ambient isotopy and approximates it at the scale ε^2. The order of magnitude ε^2 with an input $\mathcal{X}$ of scale ε is remarkable. Another instance of this phenomenon is present in [Cla06] in codimension 1. We will show that this ε^2 provides the minimax rate of approximation when dealing with random samples. Therefore, it can be thought of as optimal.

Theorem IV.10 suffers two major imperfections. First, it requires the knowledge of the tangent spaces at each sample point — since $\omega_* = \omega_{*\mathcal{X},T}$ — and it is no longer usable if tangent spaces are only known up to some error. Second, the points are assumed to lie exactly on the submanifold M, and no noise is allowed. The analysis of $\mathrm{Del}^{\omega_*}(\mathcal{X}, T)$ is sophisticated [BG14]. Rather than redo the whole study with milder assumptions, we tackle this question with an approximation theory approach (Theorem IV.11). Instead of studying if $\mathrm{Del}^{\omega_*}(\mathcal{X}', T')$ is stable when $\mathcal{X}'$ lies close to M and T' close to T, we examine what $\mathrm{Del}^{\omega_*}(\mathcal{X}', T')$ actually reconstructs, as detailed in Section IV.4.

IV.3.3 On the Sparsity Assumption

In Theorem IV.10, $\mathcal{X}$ is assumed to be dense enough so that it covers all the areas of M. It is also supposed to be sparse at the same scale as the density parameter ε. Indeed, arbitrarily accumulated points would generate non-uniformity and instability for $\mathrm{Del}^{\omega_*}(\mathcal{X}, T)$ [BGO09, BG14]. At this stage, we emphasize that the construction of a $(\varepsilon, 2\varepsilon)$-net can be carried out given an ε-dense sample. Given an ε-dense sample $\mathcal{X}$, the *farthest point sampling* algorithm prunes $\mathcal{X}$ and outputs an $(\varepsilon, 2\varepsilon)$-net $\mathcal{Y} \subset \mathcal{X}$ of M as follows. Initialize at $\mathcal{Y} = \{p_1\} \subset \mathcal{X}$, and while $\max_{p\in\mathcal{X}} d(p, \mathcal{Y}) > \varepsilon$, add to $\mathcal{Y}$ the farthest point to $\mathcal{Y}$ in $\mathcal{X}$, that is, $\mathcal{Y} \leftarrow \mathcal{Y} \cup \{\operatorname{argmax}_{p\in\mathcal{X}} d(p, \mathcal{Y})\}$. The output $\mathcal{Y}$ is ε-sparse and satisfies $d_H(\mathcal{X}, \mathcal{Y}) \leq \varepsilon$, so it is a $(\varepsilon, 2\varepsilon)$-net of M. Therefore, up to the multiplicative constant 2, sparsifying $\mathcal{X}$ at scale ε will not deteriorate its density property. Then, we can run the farthest point sampling algorithm to preprocess the data, so that the obtained point cloud is a net.

IV.4 Stability Result

IV.4.1 Interpolation Theorem

As mentioned above, if the data do not lie exactly on M and if we do not have the exact knowledge of the tangent spaces, Theorem IV.10 does not apply. To bypass this issue, we interpolate the data with another submanifold M' satisfying good properties, as stated in the following result.

Theorem IV.11 (Interpolation). *Let $M \in \mathcal{C}^2_{\tau_{min}}$. Let $\mathcal{X} = \{p_1, \dots, p_q\} \subset \mathbb{R}^D$ be a finite point cloud and $\tilde{T} = \left\{\tilde{T}_1, \dots, \tilde{T}_q\right\}$ be a family of d-dimensional linear subspaces of $\mathbb{R}^D$. For $\theta \leq \pi/64$ and $18\eta < \delta \leq \tau_{min}$, assume that*

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

Then, there exist a universal constant $c_0 \leq 285$ and a smooth submanifold $M' \subset \mathbb{R}^D$ such that

1. *$\mathcal{X} \subset M'$,*
2. *$\tau_{M'} \geq \left(1 - c_0\left(\frac{\eta}{\delta} + \theta\right)\frac{\tau_{min}}{\delta}\right)\tau_{min}$,*
3. *$T_{p_j}M' = \tilde{T}_j$ for all $1 \leq j \leq q$,*
4. *$d_H(M, M') \leq \delta\theta + \eta$,*
5. *M and M' are ambient isotopic.*

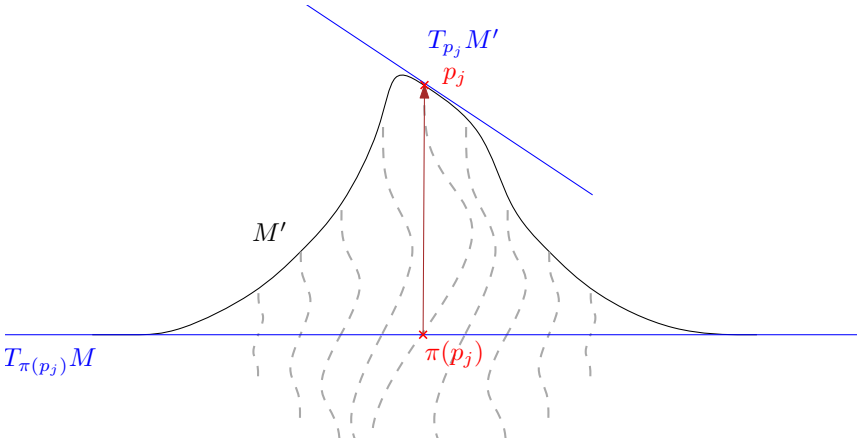


Figure IV.3 – An instance of the interpolating submanifold M'. Dashed lines correspond to the image of vertical lines by the ambient diffeomorphism Φ defining $M' = \Phi(M)$.

Theorem IV.11 fits a submanifold M' to noisy points and perturbed tangent spaces with no change of topology and a controlled reach loss. We will use M' as a proxy for M. Indeed, if $\tilde{T}_1, \dots, \tilde{T}_q$ are estimated tangent spaces at the noisy base points $p_1, \dots, p_q$, M' has the virtue of being reconstructed by $\mathrm{Del}^{\omega_*}(\mathcal{X}, \tilde{T})$ from Theorem IV.10. Since M' is topologically and geometrically close to M, we conclude that M is reconstructed as well by transitivity. In other words, Theorem IV.11 allows to consider a noisy sample with estimated tangent spaces as an exact sample with exact tangent spaces. M' is built pushing

and rotating M towards the p_j's locally along the vector $(p_j - \pi(p_j))$, as illustrated in Figure IV.3. Since the construction is quite general and may be applied in various settings, let us provide an outline of the construction.

Let $\phi(x) = \exp(\frac{\|x\|^2}{\|x\|^2-1})\mathbb{1}_{\|x\|^2<1}$. ϕ is smooth and satisfies $\phi(0) = 1$, $\|\phi\|_\infty \leq 1$ and $d_0\phi = 0$.

For $j = 1, \ldots, q$, Proposition III.29 asserts that there exists a rotation R_j of $\mathbb{R}^D$ mapping $T_{\pi_M(p_j)}M$ onto $\tilde{T}_j$ that satisfies $\|R_j - I_D\|_{op} \leq 2\sin(\theta/2) \leq \theta$. For $\ell > 0$ to be chosen later, and all $a \in \mathbb{R}^D$, let us define $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ by

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

Φ is designed to map $\pi(p_j)$ onto p_j with $d_{\pi(p_j)}\Phi = R_j$. Roughly speaking, in balls of radii ℓ around each $\pi(p_j)$, Φ shifts the points in the direction $p_j - \pi(p_j)$ and rotates it around $\pi(p_j)$. Off these balls, Φ is the identity map. To guarantee smoothness, the shifting and the rotation are modulated by the kernel ϕ, as $\|a - \pi(p_j)\|$ increases. Notice that $d_a\psi_j = (R_j - I_D)$ and $\|\psi_j(a)\| \leq \ell\theta + \eta$ whenever $\phi\left(\frac{a-\pi(p_j)}{\ell}\right) \neq 0$. Defining $M' = \Phi(M)$, the facts that M' fits to $\mathcal{X}$ and $\tilde{T}$ and is Hausdorff-close to M follow by construction. Moreover, Theorem 4.19 of [Fed59] (reproduced as Lemma B.1 in this paper) states that the reach is stable with respect to $\mathcal{C}^2$-diffeomorphisms of the ambient space. The estimate on $\tau_{M'}$ relies on the following lemma stating differentials estimates on Φ.

Lemma IV.12. *There exist universal constants $C_1 \leq 7/2$ and $C_2 \leq 28$ such that if $6\eta < \ell \leq \delta/3$ and $\theta \leq \pi/64$, $\Phi : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ is a global $\mathcal{C}^\infty$-diffeomorphism. In addition, for all a in $\mathbb{R}^D$,*

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

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

Remark IV.13. *Changing slightly the construction of M', one can also build it such that the curvature tensor at each p_j corresponds to that of M at $\pi(p_j)$. For this purpose it suffices to take a localizing function ϕ identically equal to 1 in a neighborhood of 0. This additional condition would impact the universal constant c_0 appearing in Theorem IV.11.*

IV.4.2 Stability of the Tangential Delaunay Complex

Theorem IV.11 shows that even in the presence of noisy sample points at distance η from M, and with the knowledge of the tangent spaces up to some angle θ, it is still possible to apply Theorem IV.10 to some virtual submanifold M'. Denoting $\tilde{M} = \text{Del}^{\omega_*}(\mathcal{X}, \tilde{T})$, since $d_H(M, \tilde{M}) \leq d_H(M, M') + d_H(M', \tilde{M})$ and since the ambient isotopy relation is transitive, $M \cong M' \cong \tilde{M}$. We get the following result as a straightforward combination of Theorem IV.10 and Theorem IV.11.

Theorem IV.14 (Stability of the Tangential Delaunay Complex)**.** *There exists $\varepsilon_d > 0$ such that for all $\varepsilon \leq \varepsilon_d\tau_{min}$ and all $M \in \mathcal{C}^2_{\tau_{min}}$, the following holds. Let $\mathcal{X} \subset \mathbb{R}^D$ finite point cloud and $\tilde{T} = \left\{\tilde{T}_p\right\}_{p\in\mathcal{X}}$ be a family of d-dimensional linear subspaces of $\mathbb{R}^D$ such that*

- $\max_{p\in\mathcal{X}} d(p,M) \leq \eta$,
- $\mathcal{X}$ *is* ε*-sparse,*
- $\max_{p\in\mathcal{X}} \angle(T_{\pi_M(p)}M, \tilde{T}_p) \leq \sin\theta$,
- $\max_{x\in M} d(x,\mathcal{X}) \leq 2\varepsilon$.

If $\theta \leq \varepsilon/(1140\tau_{min})$ *and* $\eta \leq \varepsilon^2/(1140\tau_{min})$*, then,*

- $d_H\left(M, \mathrm{Del}^{\omega_*}(\mathcal{X}, \tilde{T})\right) \leq C_d\varepsilon^2/\tau_{min}$,
- M *and* $\mathrm{Del}^{\omega_*}(\mathcal{X}, \tilde{T})$ *are ambient isotopic.*

Indeed, applying the reconstruction algorithm of Theorem IV.10 even in the presence of noise and uncertainty on the tangent spaces actually recovers the submanifold M' built in Theorem IV.11. M' is isotopic to M and the quality of the approximation of M is at most impacted by the term $d_H(M,M') \leq \varepsilon\theta + \eta$. The lower bound on $\tau_{M'}$ is crucial, as constants appearing in Theorem IV.10 are not bounded for arbitrarily small reach.

It is worth noting that no extra analysis of the Tangential Delaunay Complex was needed to derive its stability. The argument is global, constructive, and may be applied to other reconstruction methods taking tangent spaces as input. For instance, a stability result similar to Theorem IV.14 could be derived readily for the so-called *cocone* complex [CDR05] using the interpolating submanifold of Theorem IV.11.

IV.5 Tangent Space Estimation and Decluttering Procedure

IV.5.1 Additive Noise Case

We now focus on the estimation of tangent spaces in the model with additive noise $\mathcal{P}^2_{\tau_{min},\sigma}(f_{min}, f_{max})$. The proposed method is similar to that of [ACLZ17, MMS16]. A point $p \in M$ being fixed, T_pM is the best local d-dimensional linear approximation of M at p. Performing a Local Principal Component Analysis (PCA) in a neighborhood of p is likely to recover the main directions spanned by M at p, and therefore yield a good approximation of T_pM. For $j = 1,\dots,n$ and $h > 0$ to be chosen later, define the local covariance matrix at X_j by

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

where $\bar{X}_j = \frac{1}{N_j}\sum_{i\neq j} X_i \mathbb{1}_{\mathcal{B}(X_j,h)}(X_i)$ is the barycenter of sample points contained in the ball $\mathcal{B}(X_j,h)$, and $N_j = |\mathcal{B}(X_j,h)\cap\mathbb{X}_n|$. Let us emphasize the fact that the normalization $1/(n-1)$ in the definition of $\hat{\Sigma}_j$ stands for technical convenience. In fact, any other normalization would yield the same guarantees on tangent spaces since only the principal directions of $\hat{\Sigma}_j$ play a role. Set $\hat{T}_j(h)$ to be the linear space spanned by the d eigenvectors associated with the d largest eigenvalues of $\hat{\Sigma}_j(h)$. Computing a basis of $\hat{T}_j(h)$ can be performed naively using a singular value decomposition of the full matrix $\hat{\Sigma}_j(h)$, although fast PCA algorithms [SP07] may lessen the computational dependence on the ambient dimension. We also denote by $\texttt{TSE}(.,h)$ the function that maps any vector of points to the vector of their estimated tangent spaces, with

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

Proposition IV.15. *Set* $h = \left(c_{d,f_{min},f_{max}} \frac{\log n}{n-1}\right)^{1/d}$ *for* $c_{d,f_{min},f_{max}}$ *large enough. Assume that* $\sigma/h \leq 1/4$. *Then for* n *large enough, for all* $Q \in \mathcal{P}^2_{\tau_{min},\sigma}(f_{min}, f_{max})$,

$$\max_{1\leq j\leq n} \angle\left(T_{\pi_M(X_j)}M, \hat{T}_j(h)\right) \leq C_{d,f_{min},f_{max}}\left(\frac{h}{\tau_{min}} + \frac{\sigma}{h}\right),$$

with probability larger than $1 - 4\left(\frac{1}{n}\right)^{\frac{2}{d}}$.

An important feature given by Proposition IV.15 is that the statistical error of our tangent space estimation procedure does not depend on the ambient dimension D. The intuition behind Proposition IV.15 is the following: if we assume that the true tangent space $T_{X_j}M$ is spanned by the first d vectors of the canonical basis, we can decompose $\hat{\Sigma}_j$ as

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

where $\hat{R}$ comes from the curvature of the submanifold along with the additive noise, and is of order $N_j(h)(h^3/(\tau_{min}(n-1)) + h\sigma) \lesssim h^{d+2}(h/\tau_{min} + \sigma/h)$, provided that h is roughly smaller than $(\log(n)/(n-1))^{1/d}$. On the other hand, for a bandwidth h of order $(\log(n)/(n-1))^{1/d}$, $\hat{A}_j(h)$ can be proved (Lemma B.14) to be close to its deterministic counterpart

$$A_j(h) = \mathbb{E}\left(\left(\pi_{T_{X_j}M}(X) - \mathbb{E}\pi_{T_{X_j}M}(X)\right)\left(\pi_{T_{X_j}M}(X) - \mathbb{E}\pi_{T_{X_j}M}(X)\right)^t \mathbb{1}_{\mathcal{B}(X_j,h)}(X)\right),$$

where $\pi_{T_{X_j}M}$ denotes orthogonal projection onto $T_{X_j}M$ and expectation is taken conditionally on X_j. The bandwidth $(\log(n)/(n-1))^{1/d}$ may be thought of as the smallest radius that allows enough sample points in balls to provide an accurate estimation of the covariance matrices. Then, since $f_{min} > 0$, Lemma B.13 shows that the minimum eigenvalue of $A(h)$ is of order h^{d+2}. At last, an eigenvalue perturbation result (Proposition B.16) shows that $\hat{T}_j(h)$ must be close to $T_{X_j}M$ up to $(h^{d+3}/\tau_{min} + h^{d+1}\sigma)/(h^{d+2}) \approx h/\tau_{min} + \sigma/h$. The complete derivation is provided in Section B.5.1.

Then, it is shown in Lemma B.11, based on the results of [CGLM15], that letting $\varepsilon = c_{d,f_{min},f_{max}}(h \vee \tau_{min}\sigma/h)$ for $c_{d,f_{min},f_{max}}$ large enough, entails $\mathbb{X}_n$ is ε-dense in M with probability larger than $1 - \left(\frac{1}{n}\right)^{2/d}$. Since $\mathbb{X}_n$ may not be sparse at the scale ε, and for the stability reasons described in Section IV.3, we sparsify it with the farthest point sampling algorithm (Section IV.3.3) with scale parameter ε. Let $\mathbb{Y}_n$ denote the output of the algorithm. If $\sigma \leq h/4$, and $c_{d,f_{min},f_{max}}$ is large enough, we have the following.

Corollary IV.16. *With the above notation, for* n *large enough, with probability at least* $1 - 5\left(\frac{1}{n}\right)^{2/d}$,

- $\max_{X_j\in\mathbb{Y}_n} d(X_j, M) \leq \frac{\varepsilon^2}{1140\tau_{min}}$,
- $\mathbb{Y}_n$ *is* ε*-sparse,*
- $\max_{X_j\in\mathbb{Y}_n} \angle(T_{\pi_M(X_j)}M, \hat{T}_j(h)) \leq \frac{\varepsilon}{2280\tau_{min}}$,
- $\max_{x\in M} d(x, \mathbb{Y}_n) \leq 2\varepsilon$.

In other words, the previous result shows that $\mathbb{Y}_n$ satisfies the assumptions of Theorem IV.14 with high probability. We may then define $\hat{M}_{\texttt{TDC}}$ to be the Tangential Delaunay Complex computed on $\mathbb{Y}_n$ and the collection of estimated tangent spaces $\texttt{TSE}(\mathbb{X}_n, h)_{\mathbb{Y}_n}$, that is elements of $\texttt{TSE}(\mathbb{X}_n, h)$ corresponding to elements of $\mathbb{Y}_n$, where h is the bandwidth defined in Proposition IV.15.

Definition IV.17. *With the above notation, define* $\hat{M}_{\texttt{TDC}} = \mathrm{Del}^{\omega_*}\left(\mathbb{Y}_n, \texttt{TSE}(\mathbb{X}_n, h)_{\mathbb{Y}_n}\right)$.

Combining Theorem IV.14 and Corollary IV.16, it is clear that $\hat{M}_{\texttt{TDC}}$ satisfies Theorem IV.7.

IV.5.2 Clutter Noise Case

Let us now focus on the model with outliers $\mathcal{P}^2_{\tau_{min},\beta}(f_{min}, f_{max})$. We address problem of decluttering the sample $\mathbb{X}_n$, that is, to remove outliers. We follow ideas from [GPPVW12a]. To distinguish whether X_j is an outlier or belongs to M, we notice again that points drawn from M approximately lie on a low dimensional structure. On the other hand, the neighborhood points of an outlier drawn far away from M should typically be distributed in an isotropic way. Let $k_1, k_2, h > 0$, $x \in \mathbb{R}^D$ and $T \subset \mathbb{R}^D$ a d-dimensional linear subspace. The *slab* at x in the direction T is the set $S(x,T,h) = \{x\} \oplus \mathcal{B}_T(0, k_1 h) \oplus \mathcal{B}_{T^\perp}(0, k_2 h^2) \subset \mathbb{R}^D$, where $\oplus$ denotes the Minkovski sum, and $\mathcal{B}_T, \mathcal{B}_{T^\perp}$ are the Euclidean balls in T and $T^\perp$ respectively.

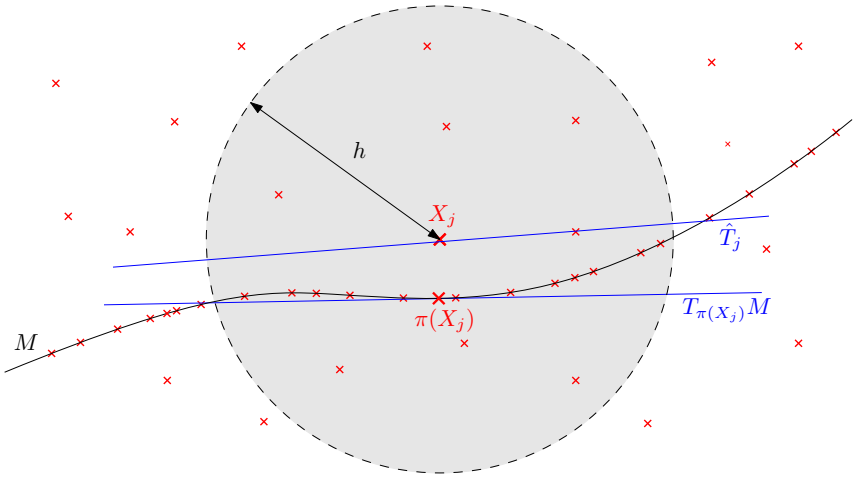


Figure IV.4 – Local PCA at an outlier point $X_j \in \mathbb{X}_n$.

Following notation of Section IV.2.1, for $P \in \mathcal{P}^2_{\tau_{min},\beta}(f_{min}, f_{max})$, let us write $P = \beta Q + (1-\beta)U_{\mathcal{B}_0}$. For h small enough, by definition of the slabs, $U_{\mathcal{B}_0}\left(S(x, T_{\pi(x)}M, h)\right) \asymp (k_1 h)^d (k_2 h^2)^{D-d} \asymp h^{2D-d}$. Furthermore, Figure IV.5 indicates that for k_1 and k_2 small enough, $Q\left(S(x, T_{\pi(x)}M, h)\right) \asymp Vol\left(S(x, T_{\pi(x)}M, h) \cap M\right) \asymp h^d$ if $d(x,M) \leq h^2$, and $Q\left(S(x, T_{\pi(x)}M, h)\right) = 0$ if $d(x,M) > h^2$. Coming back to $P = \beta Q + (1-\beta)U_{\mathcal{B}_0}$, we roughly get

$$\begin{aligned} P\left(S(x, T_{\pi(x)}M, h)\right) &\asymp \beta h^d + (1-\beta)h^{2D-d} \asymp h^d \quad \text{if} \quad d(x,M) \leq h^2, \\ P\left(S(x, T_{\pi(x)}M, h)\right) &\asymp 0 + (1-\beta)h^{2D-d} \asymp h^{2D-d} \quad \text{if} \quad d(x,M) > h^2, \end{aligned}$$

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

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

$$p_j \in \texttt{SD}_t(\mathcal{X}, T_{\mathcal{X}}, h) \quad \Leftrightarrow \quad |S(p_j, T_j, h) \cap \mathcal{X}| \geq t(n-1)h^d,$$

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

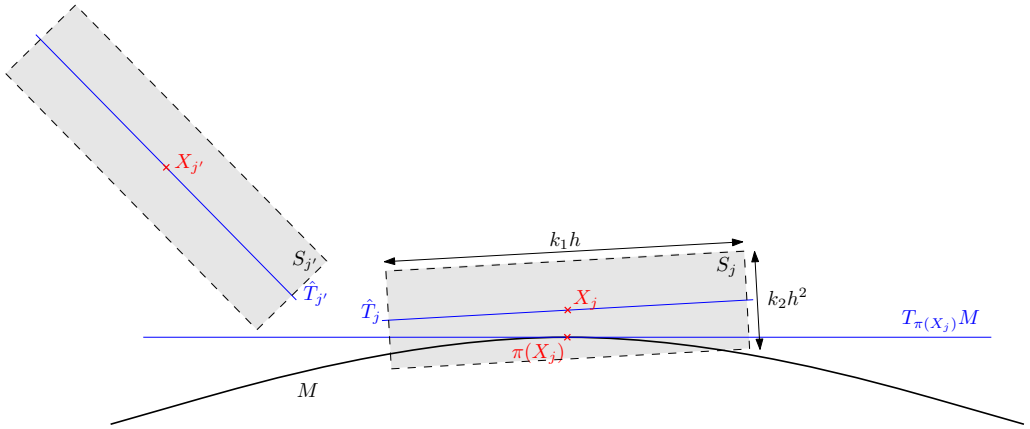


Figure IV.5 – The slab $S(X_j, \hat{T}_j, h)$ is centered at X_j and has size $k_1 h$ in the d directions spanned by $\hat{T}_j$, and size $k_2 h^2$ in the $D-d$ directions normal to $\hat{T}_j$.

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

Lemma IV.18. *Let $K > 0$ be fixed. There exist constants $k_1(K)$ and $k_2(\tau_{min}, K)$ such that for every $h \leq 1$ and x in $\mathbb{R}^D$, $S(x,T,h) \subset \mathcal{B}(x, h/2)$. Moreover, for every $h \leq h_+ \wedge 1$ we have*

$$\left.\begin{array}{c} h/\sqrt{2} \geq d(x,M) \geq h^2/\tau_{min} \\ \angle\left(T_{\pi_M(x)}M, T\right) \leq Kh/\tau_{min} \end{array}\right\} \Rightarrow S(x,T,h) \subset S'(x, T_{\pi_M(x)}M, h),$$

where $S'(x, T_{\pi_M(x)}M, h)$ is a larger slab with parameters $k_1'(\tau_{min}, K)$ and $k_2'(\tau_{min}, K)$, and satisfies $S'(x, T_{\pi_M(x)}M, h) \cap M = \emptyset$. In addition, there exists $k_3(\tau_{min}, K)$ such that for all x and y are in M,

$$\left.\begin{array}{c} \angle(T_xM, T) \leq Kh/\tau_{min} \\ \|x-y\| \leq k_3 h \end{array}\right\} \Rightarrow \quad y \in S(x,T,h).$$

Possible values for k_1 and k_2 are, respectively, $\frac{1}{16(K\vee 1)}$ and $\frac{1}{16(\tau_{min}\vee K\vee 1)}$, and k_3 can be taken as $k_1 \wedge \frac{\tau_{min} k_2}{1+2K}$.

The proof of Lemma IV.18, mentioned in [GPPVW12a], follows from elementary geometry, combined with the definition of the reach and Proposition B.3.

Roughly, Lemma IV.18 states that the decluttering performance is of order the square of the tangent space precision, hence will be closely related to the performance of the tangent space estimation procedure `TSE`. Unfortunately, a direct application of `TSE` to the corrupted sample $\mathbb{X}_n$ leads to slightly worse precision bounds, in terms of angle deviation. Typically, the angle deviation would be of order $n^{-1/(d+1)}$. However, this precision is enough to remove outliers points which are at distance at least $n^{-2/(d+1)}$ from M. Then running our `TSE` on this refined sample $\texttt{SD}_t(\mathbb{X}_n, \texttt{TSE}(\mathbb{X}_n), n^{-1/(d+1)})$ leads to better angle deviation rates, hence better decluttering performance, and so on.

Let us introduce an iterative decluttering procedure in a more formal way. We choose the initial bandwidth $h_0 = \left(c_{d,f_{min},f_{max},\tau_{min}} \frac{\log n}{\beta(n-1)}\right)^{\gamma_0}$, with $\gamma_0 = 1/(d+1)$, and define

the first set $\mathbb{X}^{(-1)} = \mathbb{X}_n$ as the whole sample. We then proceed recursively, setting $h_{k+1} = \left(c_{d,f_{min},f_{max},\tau_{min}} \frac{\log n}{\beta(n-1)}\right)^{\gamma_{k+1}}$, with γ_{k+1} satisfying $\gamma_{k+1} = (2\gamma_k + 1)/(d+2)$. This recursion formula is driven by the optimization of a trade-off between imprecision terms in tangent space estimation, as may be seen from (B.19). An elementary calculation shows that

$$\gamma_k = \frac{1}{d} - \frac{1}{d(d+1)} \left(\frac{2}{d+2}\right)^k.$$

With this updated bandwidth we define

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

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

Proposition IV.19. *In the clutter noise model, for t, $c_{d,f_{min},f_{max},\tau_{min}}$ and n large enough, k_1 and k_2 small enough, the following properties hold with probability larger than $1-7\left(\frac{1}{n}\right)^{2/d}$ for all $k \geq 0$.*

***Initialization**:*

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

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

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

***Iterations**:*

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

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

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

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

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

According to Proposition IV.19, the decluttering procedure removes no data point on M with high probability. In other words, $\mathbb{X}^{(k_\delta)} \cap M = \mathbb{X}_n \cap M$, and as a consequence, $\max_{x \in M} d(x, \mathbb{X}^{(k_\delta)}) \leq c_{d,f_{min}} \left(\frac{\log n}{\beta n}\right)^{1/d} \ll h_{k_\delta}$ with high probability (see Lemma B.11). As a consequence, we obtain the following.

Corollary IV.20. *With the above notation, for n large enough, with probability larger than $1 - 8\left(\frac{1}{n}\right)^{2/d}$,*

- $\max_{X_j \in \mathbb{Y}_n^\delta} d(X_j, M) \leq \frac{\varepsilon^2}{1140\tau_{min}}$,
- $\mathbb{Y}_n^\delta$ *is* ε*-sparse,*
- $\max_{X_j \in \mathbb{Y}_n^\delta} \angle(T_{\pi_M(X_j)}M, \hat{T}_j^\delta) \leq \frac{\varepsilon}{2280\tau_{min}}$,
- $\max_{x \in M} d(x, \mathbb{Y}_n^\delta) \leq 2\varepsilon$.

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

Definition IV.21. *With the above notation, define* $\hat{M}_{\texttt{TDC}\delta} = \mathrm{Del}^{\omega_*}\left(\mathbb{Y}_n^\delta, \hat{T}^\delta\right)$.

Combining Theorem IV.14 and Corollary IV.20, it is clear that $\hat{M}_{\texttt{TDC}\delta}$ satisfies Theorem IV.8.

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

Proposition IV.22. *Under the assumptions of Corollary IV.20, for every* $X_j \in X^{(\hat{k}+1)}$, *we have*

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

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

Definition IV.23. *With the above notation, define* $\hat{M}_{\texttt{TDC}+} = \mathrm{Del}^{\omega_*}\left(\mathbb{Y}_n^+, T^+\right)$.

In turn, Proposition IV.22 implies that $\hat{M}_{\texttt{TDC}+}$ satisfies Theorem IV.9.

IV.6 Conclusion

In this work, we gave results on explicit manifold reconstruction with simplicial complexes. We built estimators $\hat{M}_{\texttt{TDC}}$, $\hat{M}_{\texttt{TDC}\delta}$ and $\hat{M}_{\texttt{TDC}+}$ in two statistical models. We proved minimax rates of convergence for the Hausdorff distance and consistency results for ambient isotopic reconstruction. Since $\hat{M}_{\texttt{TDC}}$ is minimax optimal in the additive noise model for σ small, and uses the Tangential Delaunay Complex of [BG14], the latter is proved to be optimal. Moreover, rates of [GPPVW12a] are proved to be achievable with simplicial complexes that are computable using existing algorithms. To prove the stability of the Tangential Delaunay Complex, a generic interpolation result was derived. In the process, a tangent space estimation procedure and a decluttering method both based on local PCA were studied.

In the model with outliers, the proposed reconstruction method achieves a rate of convergence that can be as close as desired to the minimax rate of convergence, depending on the number of iterations of the decluttering procedure. Though this procedure seems to be well adapted to our reconstuction scheme — which is based on tangent spaces estimation — we believe that it could be of interest in the context of other applications. Also, further investigation may be carried out to compare this decluttering procedure to existing ones [BDWW15, Don95].

As briefly mentioned below Theorem IV.7, our approach is likely to be suboptimal in cases where noise level σ is large. In such cases, with additional structure on the noise such as *centered* and *independent from the source*, other statistical procedures such

as deconvolution [GPPVW12a] could be adapted to provide vertices to the Tangential Delaunay Complex. Tangential properties of deconvolution are still to be studied.

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

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

since the precision δ requires no more than $\log(1/\delta)$ iterations of the decluttering procedure. It is likely that better complexity bounds may be obtained using more refined algorithms, such as fast PCA [SP07], that lessens the dependence on the ambient dimension D. An interesting development would be to investigate a possible precision/complexity tradeoff, as done in [ACV14] for community detection in graphs for instance.

Even though Theorem IV.11 is applied to submanifold estimation, we believe it may be applied in various settings. Beyond its statement, the way that it is used is quite general. When intermediate objects (here, tangent spaces) are used in a procedure, this kind of proxy method can provide extensions of existing results to the case where these objects are only approximated.

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

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

Appendix B

Proofs for Chapter IV

Content

B.1 Interpolation Theorem

This section is devoted to prove the interpolation results of Section IV.4.1. For sake of completeness, let us state again Lemma III.17, a stability result for the reach with respect to $\mathcal{C}^2$-diffeomorphisms.

Lemma B.1 (Theorem 4.19 in [Fed59])**.** *Let $A \subset \mathbb{R}^D$ with $\tau_A \geq \tau_{min} > 0$ and $\Phi : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ is a $\mathcal{C}^1$-diffeomorphism such that Φ,Φ^{-1}, and $d\Phi$ are Lipschitz with Lipschitz constants K,N and R respectively, then*

$$\tau_{\Phi(A)} \geq \frac{1}{(K\tau_{min}^{-1} + R)N^2}.$$

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

$$\Phi(a) = a + \sum_{j=1}^{q} \phi_\ell\left(a - \pi(p_j)\right)\psi_j(a). \tag{B.2}$$

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

Proof of Lemma IV.12. First notice that the sum appearing in (B.2) consists of at most one term. Indeed, since $\phi \equiv 0$ outside $\mathcal{B}(0,1)$, if $\phi_\ell\left(a - \pi(p_j)\right) \neq 0$ for some $j \in \{1, \dots, q\}$, then $\|a - \pi(p_j)\| \leq \ell$. Consequently, for all $i \neq j$,

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

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

Global diffeomorphism: As the sum in (B.2) is at most composed of one term, chain rule yields

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

where the last line follows from $b_1 \leq 5/2$, $6\eta \leq \ell$ and $\theta \leq \pi/64$. Therefore, $d_a\Phi$ is invertible for all $a \in \mathbb{R}^D$, and $(d_a\Phi)^{-1} = \sum_{i=0}^{\infty}\left(I_D - d_a\Phi\right)^i$. Φ is a local diffeomorphism according to the local inverse function theorem. Moreover, $\|\Phi(a)\| \to \infty$ as $\|a\| \to \infty$, so that Φ is a global $\mathcal{C}^\infty$-diffeomorphism by Hadamard-Cacciopoli theorem [DMGZ94].

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

$$\left\|d_a\Phi\right\|_{\text{op}} \leq \left\|I_D\right\|_{op} + \left\|d_a\Phi - I_D\right\|_{op} \leq 1 + (b_1+1)\theta + b_1\frac{\eta}{\ell}.$$

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

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

where the first inequality holds since $\left\|d_a\Phi - I_D\right\|_{\text{op}} < 1$, and $\|\cdot\|_{\text{op}}$ is sub-multiplicative.
(iii) Second order: Again, since the sum (B.2) includes at most one term,

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

□

Proof of Theorem IV.11. Set $\ell = \delta/3$ and $M' = \Phi(M)$.

- *Interpolation:* For all j, $p_j = \Phi(\pi(p_j)) \in M'$ by construction since $\phi_\ell(0) = 1$.

- *Tangent spaces:* Since $d_x\phi_l|_{x=0} = 0$, for all $j \in \{1, \dots, q\}$, $d_a\Phi|_{a=\pi(p_j)} = R_j$. Thus,

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

  by definition of R_j.

- *Proximity to M:* The bound on $d_H(M, M') = d_H(M, \Phi(M))$ follows from the correspondence

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

- *Isotopy:* Consider the continuous family of maps

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

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

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

$$\tau_{M'} \geq \frac{\left(1 - C_1\left(\frac{\eta}{\ell} + \theta\right)\right)^2}{\frac{1 + C_1\left(\frac{\eta}{\ell} + \theta\right)}{\tau_{min}} + C_2\left(\frac{\eta}{\ell^2} + \frac{\theta}{\ell}\right)} = \tau_{min} \cdot \frac{\left(1 - C_1\left(\frac{\eta}{\ell} + \theta\right)\right)^2}{1 + C_1\left(\frac{\eta}{\ell} + \theta\right) + C_2\left(\frac{\eta}{\ell^2} + \frac{\theta}{\ell}\right)\tau_{min}}.$$

  Now, replace ℓ by its value $\delta/3$, and write $c_1 = 3C_1 \leq 21/2 \leq 11$ and $c_2 = 3^2C_2 \leq 252$. We derive

$$\begin{aligned} \tau_{M'} &\geq \left(1 - 2c_1\left(\frac{\eta}{\delta} + \theta\right)\right)\left(1 - c_1\left(\frac{\eta}{\delta} + \theta\right) - c_2\left(\frac{\eta}{\delta^2} + \frac{\theta}{\delta}\right)\tau_{min}\right)\tau_{min} \\ &\geq \left(1 - 3c_1\left(\frac{\eta}{\delta} + \theta\right) - c_2\left(\frac{\eta}{\delta^2} + \frac{\theta}{\delta}\right)\tau_{min}\right)\tau_{min} \\ &\geq \left(1 - (3c_1 + c_2)\left(\frac{\eta}{\delta^2} + \frac{\theta}{\delta}\right)\tau_{min}\right)\tau_{min}, \end{aligned}$$

  where for the last line we used that $\delta/\tau_{min} \leq 1$. The desired lower bound follows taking $c_0 = 3c_1 + c_2 \leq 285$.

□

B.2 Some Geometric Properties under Reach Regularity Condition

B.2.1 Reach and Projection on the Submanifold

In this section we state intermediate results that connect the reach condition to orthogonal projections onto the tangent spaces. Let us start by restating (III.19) the precise way we will us it.

Proposition B.3 (Theorem 4.18 in [Fed59])**.** *For all x and y in M,*

$$\|(y-x)_\perp\| \leq \frac{\|y-x\|^2}{2\tau_{min}},$$

where $(y-x)_\perp$ denotes the projection of $y-x$ onto $T_xM^\perp$.

From Proposition B.3 we may deduce the following property about trace of Euclidean balls on M.

Proposition B.4. *Let $x \in \mathbb{R}^D$ be such that $d(x,M) = \Delta \leq h \leq \frac{\tau_{min}}{8}$, and let y denote $\pi(x)$. Then,*

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

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

Proof of Proposition B.4. Let z be in $M \cap \mathcal{B}(x,h)$, and denote by δ the quantity $\|z-y\|$. We may write

$$\|z-x\|^2 = \delta^2 + \Delta^2 + 2\langle z-y, y-x\rangle, \tag{B.5}$$

hence $\delta^2 \leq h^2 - \Delta^2 - 2\langle z-y, y-x\rangle$. Denote, for u in $\mathbb{R}^D$, by $u_\perp$ its projection onto $T_yM^\perp$. Since $\langle z-y, y-x\rangle = \langle (z-y)_\perp, y-x\rangle$, Proposition B.3 ensures that

$$\delta^2\left(1 - \frac{\Delta}{\tau_{min}}\right) \leq r_h^2.$$

Since $\Delta \leq h \leq \tau_{min}/8$, it comes $\delta^2 \leq (1 + 2\frac{\Delta}{\tau_{min}}) r_h^2$. On the other hand, (B.5) and Proposition B.3 also yield

$$\|z-x\|^2 \leq \delta^2\left(1 + \frac{\Delta}{\tau_{min}}\right) + \Delta^2.$$

Hence, if $\delta^2 \leq \left(1 - \frac{\Delta}{\tau_{min}}\right) r_h^2$, we have

$$\|z-x\|^2 \leq r_h^2 + \Delta^2 = h^2.$$

□

Also, the following consequence of Proposition B.3 will be of particular use in the decluttering procedure.

Proposition B.6. *Let h and h_k be bandwidths satisfying $h_k^2/\tau_{min} \leq h \leq h_k$. Let x be such that $d(x,M) \leq h/\sqrt{2}$ and $\pi_M(x) = 0$, and let z be such that $\|z-x\| \leq h$ and $d(z,M) \leq h_k^2/\tau_{min}$. Then*

$$\|z_\perp\| \leq \frac{6h_k^2}{\tau_{min}},$$

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

Proof of Proposition B.6. Let y denote $\pi_M(z)$. A triangle inequality yields $\|y\| \leq \|y - z\| + \|z - x\| + \|x\| \leq h_k^2/\tau_{min} + (1 + 1/\sqrt{2})h \leq 3h_k$. Proposition B.3 ensures that $\|y_\perp\| \leq \|y\|^2/(2\tau_{min}) \leq (9h_k^2)/(2\tau_{min})$. Since $\|z_\perp\| \leq \|y_\perp\| + h_k^2/\tau_{min}$, we have $\|z_\perp\| \leq 6h_k^2/\tau_{min}$. □

At last, let us prove Lemma IV.18, that gives properties of intersections of ambient slabs with M.

Proof. (Proof of Lemma IV.18) Set $k_1 = \frac{1}{16(K \vee 1)}$, $k_2 = \frac{1}{16(K \vee \tau_{min} \vee 1)}$, and $k_3 = k_1 \wedge \frac{\tau_{min} k_2}{1+2K} \wedge 1$. For all $h > 0$, and $z \in S(x, T, h)$, triangle inequality yields $\|z - x\| \leq \|\pi_T(z - x)\| + \|\pi_{T^\perp}(z - x)\| \leq (k_1 + k_2 h)h$. Since $h \leq 1$ and $k_1 + k_2 \leq 1/2$, we get $z \in \mathcal{B}(x, h/2)$.

Now, suppose that $h/\sqrt{2} \geq d(x, M) \geq h^2/\tau_{min}$ and $\angle\left(T_{\pi(x)}M, T\right) \leq Kh/\tau_{min}$. For short we write $T_0 = T_{\pi(x)}M$. Let $z \in S(x, T, h)$, since $h \leq 1$, it comes

$$\|\pi_{T_0}(z - x)\| \leq \|z - x\| \leq (k_1 + k_2)h = k_1' h,$$

with $k_1' = k_1 + k_2$. On the other hand

$$\|\pi_{T_0^\perp}(z - x)\| \leq \|\pi_{T_0^\perp}\pi_T(z - x)\| + \|\pi_{T_0^\perp}\pi_{T^\perp}(z - x)\| \leq (Kh/\tau_{min})(k_1 h) + k_2 h^2 = k_2' h^2,$$

with $k_2' = k_1 K/\tau_{min} + k_2$. Hence $S(x, T, h) \subset S'(x, T_0, h)$, for the constants k_1' and k_2' defined above. It remains to prove that $S'(x, T_0, h) \cap M = \emptyset$. To see this, let $z \in S'(x, T_0, h)$, and $y = \pi(x)$. Since $k_1' + k_2' \leq 1/4$, we have $\|y - z\| \leq \|y - x\| + \|x - z\| \leq h(1/\sqrt{2} + 1/4)$. For the normal part, we may write

$$\|\pi_{T_0^\perp}(z - y)\| \geq \|\pi_{T_0^\perp}(y - x)\| - \|\pi_{T_0^\perp}(x - z)\| \geq h^2(1/\tau_{min} - k_2').$$

Since $k_2' \leq 1/(8\tau_{min})$, we have $\|\pi_{T_0^\perp}(z - y)\| > \|y - z\|^2/(2\tau_{min})$, hence Proposition B.3 ensures that $z \notin M$.

At last, suppose that $x \in M$ and $y \in \mathcal{B}(x, k_3 h) \cap M$. Since $k_3 \leq k_1$, we have $\|\pi_T(y - x)\| \leq k_1 h$. Next, we may write

$$\|\pi_{T^\perp}(y - x)\| \leq \|\pi_{T^\perp}\pi_{T_0}(y - x)\| + \|\pi_{T^\perp}\pi_{T_0^\perp}(y - x)\|.$$

Since $y \in M$, Proposition B.3 entails $\|\pi_{T_0^\perp}(y - x)\| \leq \|y - x\|^2/(2\tau_{min}) \leq k_3^2 h^2/(2\tau_{min})$. It comes

$$\|\pi_{T^\perp}(y - x)\| \leq \frac{h^2}{\tau_{min}}\left(k_3 K + \frac{k_3^2}{2}\right) \leq k_2 h^2.$$

Hence $y \in S(x, T, h)$. □

B.2.2 Reach and Exponential Map

In this section we state results that connect Euclidean and geodesic quantities under reach regularity condition. See also Chapter III for further details. We start with a result linking reach and principal curvatures.

Proposition B.7 (Proposition 6.1 in [NSW08])**.** *For all $x \in M$, writing II_x for the second fundamental form of M at x, for all unitary $w \in T_x M$, we have $\|II_x(w, w)\| \leq 1/\tau_{min}$.*

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

Proposition B.8 (Corollary 1.4 in [AB06a])**.** *The injectivity radius of M is at least $\pi\tau_{min}$.*

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

Proposition B.9. *For all $x, y \in M$ such that $\|x-y\| \leq \tau_{min}/4$,*

$$\|x-y\| \leq d_M(x,y) \leq \alpha \|x-y\|.$$

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

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

with $\|R(r,v)\| \leq \frac{r^2}{2\tau_{min}}$.

Proof of Proposition B.9. The first statement is a direct consequence of Proposition 6.3 in [NSW08]. Let us define $u(t) = \exp_x(tv) - \exp_x(0) - tv$ and $w(t) = \exp_x(tv)$ for all $0 \leq t \leq r$. It is clear that $u(0) = 0$ and $u'(0) = 0$. Moreover, $\|u''(t)\| = \left\| II_{w(t)}\left(w'(t), w'(t)\right)\right\| \leq 1/\tau_{min}$. Therefore, a Taylor expansion at order two gives $\|R(r,v)\| = u(r) \leq r^2/(2\tau_{min})$. Applying the first statement of the proposition gives $r \leq \alpha\|x-y\|$. □

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

Proposition B.10. *Let $x \in M$ be fixed. Denote by $J(r,v)$ the Jacobian of the volume form expressed in polar coordinates around x, for $r \leq \frac{\tau_{min}}{4}$ and v a unit vector in T_xM. In other words, if $y = \exp_x(rv)$, $d_yV = J(r,v)drdv$. Then*

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

where $c_d = 2^{-d}$ and $C_d = 2^d$. As a consequence, if $\mathcal{B}_M(x,r)$ denotes the geodesic ball of radius r centered at x, then, if $r \leq \frac{\tau_{min}}{4}$,

$$c'_d r^d \leq Vol(\mathcal{B}_M(x,r)) \leq C'_d r^d,$$

with $c'_d = c_d V_d$ and $C'_d = C_d V_d$, where V_d denotes the volume of the unit d-dimensional Euclidean ball.

Proof of Proposition B.10. Denoting $A_{r,v} = d_{rv}\exp_x$, the Area Formula [Fed69, Section 3.2.5] asserts that $J(r,v) = r^{d-1}\sqrt{\det\left(A_{r,v}^t A_{r,v}\right)}$. Note that from Proposition 6.1 in [NSW08] together with the Gauss equation [dC92, p. 130], the sectional curvatures in M are bounded by $|\kappa| \leq 2/\tau_{min}^2$. Therefore, the Rauch theorem [DVW15, Lemma 5] states that

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

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

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

Since $Vol(\mathcal{B}_M(x,r)) = \int_{s=0}^{r}\int_{v\in\mathcal{S}_{d-1}} J(s,v)dsdv$, where $\mathcal{S}_{d-1}$ denotes the unit $(d-1)$-dimensional sphere, the bounds on the volume easily follows. □

B.3 Some Technical Properties of the Statistical Model

B.3.1 Covering and Mass

Lemma B.11. *Let $Q_0 \in \mathcal{U}_M(f_{min}, f_{max})$. Then for all $p \in M$ and $r \leq \tau_{min}/4$,*

$$Q_0(\mathcal{B}(p,r)) \geq a_d f_{min} r^d,$$

where $a_d > 0$. As a consequence, for n large enough and for all $Q \in \mathcal{P}^2_{\tau_{min},\sigma}(f_{min}, f_{max})$, with probability larger that $1 - \left(\frac{1}{n}\right)^{2/d}$,

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

Similarly, for n large enough and for all $P \in \mathcal{P}^2_{\tau_{min},\beta}(f_{min}, f_{max})$, with probability larger that $1 - \left(\frac{1}{n}\right)^{2/d}$,

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

Proof of Lemma B.11. The first statement is a direct corollary of Lemma III.23. Let us now prove the second statement. By definition, sample $X_i \in \mathbb{X}_n$, that has distribution $Q \in \mathcal{P}^2_{\tau_{min},\sigma}(f_{min}, f_{max})$ can be written as $X_i = Y_i + Z_i$, with Y_i having distribution $Q_0 \in \mathcal{P}^2_{\tau_{min}}(f_{min}, f_{max})$, and $\|Z_i\| \leq \sigma$. From Lemma III.23 again, writing $\mathbb{Y}_n = \{Y_1, \ldots, Y_n\}$, for $r \leq \tau_{min}/8$ we obtain

$$\mathbb{P}_{Q_0}\left(d_H(M, \mathbb{Y}_n) > r\right) \leq \frac{4^d}{ar^d} \exp\left(-n\frac{a}{2^d} r^d\right).$$

The statement then follows using that $d_H(\mathbb{X}_n, \mathbb{Y}_n) \leq \sigma$, and setting $r = C_{d,f_{min}} \left(\frac{\log n}{n}\right)^{1/d}$ with $C^d_{d,f_{min}} \frac{a}{2^{d+1}} \geq 1 + 2/d$.

To prove the last point, notice that for all $k = 0, \ldots, n$, conditionally on the event $\{|\mathbb{X}_n \cap M| = k\}$, $\mathbb{X}_n \cap M$ has the distribution of a k-sample of Q_0. Therefore,

$$\begin{aligned}\mathbb{P}_P\left(d_H(M, \mathbb{X}_n \cap M) > r \,|\, |\mathbb{X}_n \cap M| = k\right) &= \mathbb{P}_{Q_0}\left(d_H(M, \mathbb{X}_k \cap M) > r\right) \\ &\leq \frac{4^d}{ar^d} \exp\left(-k\frac{a}{2^d} r^d\right).\end{aligned}$$

Hence,

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

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

Now we allow for some outliers. We consider a random variable X with distribution P, that can be written as $X = V(Y + Z) + (1 - V)X''$, with $\|Z\| \leq sh$, $s \leq 1/4$, such that $\mathbb{P}(V = 1) = \beta$ and V is independent from (Y, Z, X''), Y has law Q in $\mathcal{P}^2_{\tau_{min}}(f_{min}, f_{max})$, and X'' has uniform distribution on $\mathcal{B}(0, K_0)$ (recall that K_0 is defined below Lemma IV.2). Note that $s = 0$ corresponds to the clutter noise case, whereas $\beta = 1$ corresponds to the additive noise case.

For a fixed point x, let $p(x, h)$ denote $P(\mathcal{B}(x, h))$. We have $\mathbb{P}(VY \in \mathcal{B}(x, (1 - s)h)) \leq \mathbb{P}(VX \in \mathcal{B}(x, h)) \leq \mathbb{P}(VY \in \mathcal{B}(x, 2h))$. Hence we may write

$$\beta q(x, 3/4h) + (1 - \beta) q'(x, h) \leq p(x, h) \leq \beta q(x, 2h) + (1 - \beta) q'(x, h),$$

where $q(x, h) = Q(\mathcal{B}(x, h))$, and $q'(x, h) = (h/K_0)^D$. Bounds on the quantities above are to be found in the following lemma.

Lemma B.12. *There exists $h_+(\tau_{min}, \beta, f_{min}, f_{max}, d) \leq \tau_{min}/\sqrt{12d}$ such that, if $h \leq h_+$, for every x such that $d(x, M) \leq h$, we have*

- $\mathcal{B}(x, 2h) \cap M \subset \mathcal{B}(\pi_M(x), 4h) \cap M$,
- $q(x, 2h) \leq C_d f_{max} h^d$.

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

- $\mathcal{B}(\pi_M(x), h/8) \cap M \subset \mathcal{B}(x, 3h/4)$,
- $c_d f_{min} h^d \leq q(x, 3h/4)$,
- $p(x, h) \leq 2\beta q(x, 2h)$.

Proof of Lemma B.12. Set $h_1(\tau_{min}) = \tau_{min}/(16\alpha)$, and let x be such that $d(x, M) \leq h$, and $h \leq h_1$. According to Proposition B.4, $\mathcal{B}(x, 2h) \cap M \subset \mathcal{B}\left(\pi_M(x), r^+_{2h}\right) \cap M$, with $r^+_{2h} = \sqrt{(1 + 2\Delta/\tau_{min})} r_{2h} \leq 2r_{2h} \leq 4h$. According to Proposition B.9, if $y \in \mathcal{B}(\pi_M(x), 4h) \cap M$, then $d_M(\pi_M(x), y) \leq 4\alpha h \leq \tau_{min}/4$. Proposition B.10 then yields $q(x, 2h) \leq C_d f_{max} h^d$.

Now if $d(x, M) \leq h/\sqrt{2}$, $\mathcal{B}\left(\pi_M(x), r^-_{3h/4}\right) \cap M \subset \mathcal{B}(x, 3h/4) \cap M$ according to Proposition B.4, with $r^-_{3h/4} = \sqrt{(1 - \Delta/\tau_{min})} r_{3h/4} \geq r_{3h/4}/2 \geq h/8$. Since $\mathcal{B}_M(\pi_M(x), h/8) \subset \mathcal{B}(\pi_M(x), h/8) \cap M$, a direct application of Proposition B.10 entails $c_d f_{min} h^d \leq q(x, 3h/4)$.

Applying Proposition B.10 again, there exists $h_2(f_{min}, d, D, \beta, \tau_{min})$ such that if $h \leq h_1 \wedge h_2$, then for any x such that $d(x, M) \leq h/\sqrt{2}$ we have $(1 - \beta) q'(x, h) \leq \beta c_{d, f_{min}} h^d$, along with $q(x, 2h) \geq q(x, 3h/4) \geq c_{d, f_{min}} h^d$. We deduce that $p(x, h) \leq 2\beta q(x, 2h)$. Taking $h_+ = h_1 \wedge h_2 \wedge \tau_{min}/\sqrt{12d}$ leads to the result. □

B.3.2 Local Covariance Matrices

In this section we describe the shape of the local covariance matrices involved in tangent space estimation. Without loss of generality, the analysis will be conducted for $\hat{\Sigma}_1$ (at sample point X_1), abbreviated as $\hat{\Sigma}$. We further assume that $d(X_1, M) \leq h/\sqrt{2}$, $\pi_M(X_1) = 0$, and that $T_0 M$ is spanned by the d first vectors of the canonical basis of $\mathbb{R}^D$.

The two models (additive noise and clutter noise) will be treated jointly, by considering a random variable X of the form

$$X = V(Y + Z) + (1 - V)X'',$$

where $\mathbb{P}(V = 1) = \beta$ and V is independent from (Y, Z, X''), Y has distribution in $\mathcal{P}^2_{\tau_{min}, \sigma}(f_{min}, f_{max})$, $\|Z\| \leq \sigma$, and X'' has uniform law on $\mathcal{B}(0, K_0)$ (recall that K_0

is defined above Definition IV.4). For short we denote by s the quantity σ/h, and recall that we take $s \leq 1/4$, along with $h \leq h_+$ (defined in Lemma B.12).

Let $U(X_i, h)$, $i = 2, \ldots, n$, denote $\mathbb{1}_{\mathcal{B}(X_1,h)}(X_i)$, let $Y_i \in M$ and Z_i such that $X_i = Y_i + Z_i$, with $\|Z_i\| \leq sh$, and let $V_2, \ldots V_n$ denote random variables such that $V_i = 1$ if X_i is drawn from the signal distribution (see page 47). It is immediate that the $(U(X_i, h), V_i)$'s are independent and identically distributed, with distribution $(U(X, h), V)$.

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

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

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

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

Lemma B.13. *If $d(X_1, M) \leq h/\sqrt{2}$, for $h \leq h_+$, then*

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

with

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

Furthermore,

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

Proof of Lemma B.13. Let $x = y + z$ be in $B(X_1, h)$, with $y \in M$ and $\|z\| \leq sh$. Since $s \leq 1/4$, $\|y\| \leq 2h$. According to Proposition B.4 combined with Proposition B.9, we may write, for $h \leq h_+$ and y in $\mathcal{B}(X_1, 2h) \cap M$,

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

in local polar coordinates. Moreover, if $y \in \mathcal{B}(X_1, (1-s)h)$, then $x \in \mathcal{B}(X_1, h)$. Then, according to Proposition B.4, we have $\mathcal{B}(\pi_M(X_1), r^-_{3h/4}) \cap M \subset \mathcal{B}(X_1, (1-s)h) \cap M$. Let u be a unit vector in T_0M. Then $\langle u, x - m_\top(h)\rangle^2 = \langle u, rv + R(r,v) + z - m_\top(h)\rangle^2 \geq \langle u, rv - m_\top(h)\rangle^2 /2 - 3(R(r,v)+z)^2 \geq \langle u, rv - m_\top(h)\rangle^2 /2 - 6r^4/(4\tau_{min}^2) - 6s^2h^2$ according to Proposition B.9. Hence we may write

$$\begin{aligned} \langle Au, u\rangle &= \beta \int_{\mathcal{B}(X_1,h)\cap M} \langle u, rv + R(r,v) - m_\top(h)\rangle^2 J(r,v) f(r,v) dr dv \\ &\geq \beta f_{min} c_d \int_{r=0}^{r^-_{3h/4}} \int_{\mathcal{S}_{d-1}} r^{d-1} \left[\langle u, rv - m_\top(h)\rangle^2 /2 - 3r^4/(2\tau_{min}^2) - 6s^2h^2\right] dr dv, \end{aligned}$$

according to Proposition B.10 (bound on $J(r, v)$) and Proposition B.4 (the geodesic ball $\mathcal{B}_M(\pi_M(X_1), r^-_{3h/4})$ is included in the Euclidean ball $\mathcal{B}(\pi_M(X_1), r^-_{3h/4}) \subset \mathcal{B}(X_1, (1-s)h) \cap M$). Then

$$\begin{aligned} \int_{r=0}^{r^-_{3h/4}} \int_{\mathcal{S}_{d-1}} \frac{r^{d-1}\langle u, rv - m_\top(h)\rangle^2}{2} dr dv &\geq \int_{r=0}^{r^-_{3h/4}} \int_{\mathcal{S}_{d-1}} \frac{r^{d-1}\langle u, rv\rangle^2}{2} dr dv \\ &= \frac{\sigma_{d-1}}{2d} \int_{r=0}^{r^-_{3h/4}} r^{d+1} dr = \frac{\sigma_{d-1}(r^-_{3h/4})^{d+2}}{2d(d+2)}, \end{aligned}$$

where σ_{d-1} denotes the surface of the $d-1$-dimensional unit sphere. On the other hand,

$$\int_{r=0}^{r^-_{3h/4}} \int_{\mathcal{S}_{d-1}} \frac{3r^{d+3}}{2\tau_{min}^2} + 6s^2h^2r^{d-1} dr dv = \sigma_{d-1}(r^-_{3h/4})^{d+2} \left(\frac{3(r^-_{3h/4})^2}{2(d+4)\tau_{min}^2} + \frac{6s^2h^2}{d} \right).$$

Since $r^-_{3h/4} \leq h \leq h_+ \leq \tau_{min}/\sqrt{12d}$, we conclude that

$$\langle Au, u \rangle \geq \beta c_d f_{min} (r^-_{3h/4})^{d+2} \geq \beta c_d f_{min} h^{d+2},$$

since, for $d(X_1, M) \leq h/\sqrt{2}$ and $h \leq h_+$, $r^-_{3h/4} \geq r_{3h/4}/2 \geq h/8$, according to Proposition B.4.

Now, since for any $x = y + z \in \mathcal{B}(X_1, h)$, $y \in M \cap \mathcal{B}(0, 2h)$ and $\|z\| \leq sh$, we have $\|y_\perp\| \leq 2h^2/\tau_{min}$, according to Proposition B.3. Jensen's inequality yields that $\|m(h)_\perp\| \leq 2h^2/\tau_{min} + sh$ and $\|m(h)_\top\| \leq \|m(h)\| \leq 2h$. □

The following Lemma B.14 is devoted to quantify the deviations of empirical quantities such as local covariance matrices, means and number of points within balls from their deterministic counterparts. To this aim we define $N_0(h)$ and $N_1(h)$ as the number of points drawn from respectively noise and signal in $\mathcal{B}(X_1, h) \cap M$, namely

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

Lemma B.14. *Recall that* $h_0 = \left(\kappa \frac{\log n}{\beta(n-1)}\right)^{1/(d+1)}$ *(as defined page 58), and* $h_\infty = h_0^{(d+1)/d}$, *for* κ *to be fixed later.*

If $h_0 \leq h_+$ *and* $d(X_1, M) \leq h_+/\sqrt{2}$, *then, with probability larger than* $1 - 4\left(\frac{1}{n}\right)^{2/d+1}$, *the following inequalities hold, for all* $h \leq h_0$.

$$\begin{array}{rcl} \frac{N_0(h)}{n-1} & \leq & 2(1-\beta)q'(h) + \frac{10(2+2/d)\log n}{n-1}, \\ \frac{N_1(h)}{n-1} & \leq & 2\beta q(2h) + \frac{10(2+2/d)\log n}{n-1}. \end{array}$$

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

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

Proof of Lemma B.14. The first two inequalities are straightforward applications of Theorem 5.1 in [BBL05]. The proofs of the two last results are detailed below. They are based on Talagrand-Bousquet's inequality (see, e.g., Theorem 2.3 in [Bou02]) combined with the so-called peeling device.

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

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

for $h_- \leq h \leq h_0$, T a $d \times d$ matrix such that $\|T\|_{\mathcal{F}} = 1$, x in $\mathbb{R}^D$, v in $\{0,1\}$, and $\langle T, B\rangle = \text{trace}(T^t A)$, for any square matrices T and A. Now we define the weighted empirical process

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

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

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

for $h_- \leq u \leq h_0$. Since $\|f_{T,h}\|_\infty \leq sup_{x\in M}\|x - m(h)\|^2 U(x,h) \leq 4h^2$, and

$$\begin{aligned} Var(f_{T,h}(X,V)) \leq \mathbb{E}\left(\|X - m(h)\|^2 U(X,h)V\right) &\leq 16\beta h^4 \mathbb{P}(VX \in \mathcal{B}(X_1, h) \\ &\leq 16\beta h^4 \mathbb{P}(VY \in \mathcal{B}(X_1, 2h), \end{aligned}$$

for $s \leq 1/4$, a direct application of Theorem 2.3 in [Bou02] yields, with probability larger than $1 - e^{-x}$,

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

To get a bound on $\mathbb{E}Z(u)$, we introduce some independent Rademacher random variables $\sigma_2, \ldots, \sigma_n$, i.e. $\mathbb{P}(\sigma_j = 1) = \mathbb{P}(\sigma_j = -1) = 1/2$. With a slight abuse of notation, expectations with respect to the (X_i, V_i)'s and σ_i's, $i = 2, \ldots, n$, will be denoted by $\mathbb{E}_{(X,V)}$ and $\mathbb{E}_\sigma$ in what follows. According to the symmetrization principle (see, e.g., Lemma 11.4 in [BLM13]), we have

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

For a fixed sequence (X_i, V_i), $i = 2, \ldots, n$, we may write

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

Jensen's inequality ensures that

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

hence

$$\mathbb{E}_{(X,V)}E_{12} \leq 4u^2\sqrt{\beta(n-1)q(2u)}.$$

For the term E_{11}, note that, when $(X_i, V_i)_{i=2,\dots,n}$ is fixed,

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

is in fact a supremum of at most $N_1(u)$ processes. According to the bounded difference inequality (see, e.g., Theorem 6.2 of [BLM13]), each of these processes is subGaussian with variance bounded by $16h^4N_1(u)$ (see Theorem 2.1 of [BLM13]). Hence a maximal inequality for subGaussian random variables (see Section 2.5, p.31, of [BLM13]) ensures that

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

Hence $\mathbb{E}_{(X,V)}E_{11} \leq 4h^2\sqrt{2\beta(n-1)q(2u)\log(n-1)}$. E_2 may also be decomposed as

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

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

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

Similarly, we may write

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

At last, we may decompose E_4 as

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

using the same argument. Combining all these terms leads to

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

hence we get

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

To derive a bound on the weighted process Z, we make use of the so-called peeling device (see, e.g., Section 13.7, p.387, of [BLM13]). Set $p = \lceil \log(h_0/h_\infty)\rceil \leq 1+\log(h_0/h_\infty)$, so that $e^{-p}h_0 \leq h_-$. According to Lemma B.12, if I_j denotes the slice $[e^{-j}h_0, e^{-(j-1)}h_0] \cap [h_-, h_0]$, then, for every h in I_j, we have

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

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

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

Since $q(2h_{j-1}) \geq q(2h_-)$, we deduce that

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

Now, according to Lemma B.12, $\beta q(2h_-) \geq c_d\kappa\log n/(n-1)$. On the other hand, $p \leq 1+\log(h_0/h_\infty) \leq \log(\beta(n-1)/\kappa)/d \leq \log n/d$, for $\kappa \geq 1$. For n large enough, taking $x = (1+2/d)\log n$ in the previous inequality, we get

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

The last concentration inequality of Lemma B.14 may be derived the same way, considering the functions

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

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

B.3.3 Decluttering Rate

In this section we prove that, if the angle between tangent spaces is of order h, then we can distinguish between outliers and signal at order h^2. We recall that the slab $S(x,T,h)$ is the set of points y such that $\|\pi_T(y-x)\| \leq k_1 h$ and $\|\pi_{T^\perp}(y-x)\| \leq k_2 h^2$, k_1 and k_2 defined in Lemma IV.18, and where π_T denotes the orthogonal projection onto T.

Lemma B.15. *Recall that $h_0 = \left(\kappa \frac{\log n}{\beta(n-1)}\right)^{1/(d+1)}$, and $h_\infty = h_0^{(d+1)/d}$. Let K be fixed, and k_1, k_2 defined accordingly from Lemma IV.18. If $h_0 \leq h_+$, for κ large enough (depending on d, τ_{min} and f_{min}) and n large enough, there exists a threshold t such that, for all $h_\infty \leq h \leq h_0$, we have, with probability larger than $1 - 3\left(\frac{1}{n}\right)^{2/d+1}$,*

$$\left.\begin{array}{c} X_1 \in M \\ \angle(T, T_{X_1}M) \leq Kh/\tau_{min} \end{array}\right\} \Rightarrow |S(X_1,T,h) \cap \{X_2,\ldots,X_n\}| \geq t(n-1)h^d,$$

$$\left.\begin{array}{c} d(X_1,M) \geq h^2/\tau_{min} \\ \angle\left(T, T_{\pi(X_1)}M\right) \leq Kh/\tau_{min} \end{array}\right\} \Rightarrow |S(X_1,T,h) \cap \{X_2,\ldots,X_n\}| < t(n-1)h^d,$$

$$d(X_1,M) \geq h/\sqrt{2} \Rightarrow |S(X_1,T,h) \cap \{X_2,\ldots,X_n\}| < t(n-1)h^d.$$

Proof of Lemma B.15. Suppose that $d(X_1,M) \geq h/\sqrt{2}$. Then, according to Lemma IV.18, $S(X_1,T,h) \subset \mathcal{B}(X_1,h/2)$, with $\mathcal{B}(X_1,h/2) \cap M = \emptyset$, hence $P_n(S(X_1,T,h)) \leq P_n(\mathcal{B}(X_1,h/2))$. Theorem 5.1 in [BBL05] yields that, for all $h_\infty \leq h \leq h_0$, with probability larger than $1 - \left(\frac{1}{n}\right)^{2/d+1}$,

$$P_n(\mathcal{B}(X_1,h/2)) \leq 2P(\mathcal{B}(X_1,h/2)) + \frac{4\,(2/d+1)\log(8n)}{n-1}.$$

Since $\log(n)/(n-1) \leq \beta h^d/\kappa$, we may write

$$\begin{aligned} P_n(S(X_1,T,h)) &\leq 2Q'(\mathcal{B}(X_1,h/2)) + \frac{4\,(2/d+1)\log(8n)}{n-1} \\ &\leq 2(1-\beta)\frac{h^D}{(2K_0)^D} + \frac{4\,(2/d+1)\log(8n)}{n-1} \\ &\leq (1-\beta)C_{d,D,\tau_{min},f_{min}}h^{d+1} + \frac{4\,(2/d+1)\log(8n)}{n-1} \\ &\leq h^d\left((1-\beta)C_{d,D,\tau_{min},f_{min}}h + \frac{C_d\beta}{\kappa}\right), \end{aligned}$$

for n large enough so that $h \leq 1$.

If $h/\sqrt{2} \geq d(X_1,M) \geq h^2/\tau_{min}$ and $\angle\left(T_{\pi(X_1)}M, T\right) \leq Kh/\tau_{min}$, then Lemma IV.18 provides a big slab $S'(x,T_{\pi(x)}M,h)$ so that $S(x,T,h) \subset S'(x,T_{\pi(x)}M,h)$ and $S'(x,T_{\pi(x)}M,h) \cap M = \emptyset$. Thus, $P_n(S(x,T,h)) \leq P_n(S'(x,T_{\pi(x)}M,h))$. An other application of Theorem 5.1 in [BBL05] yields that, for all $h_\infty \leq h \leq h_0$, with probability larger than $1 - \left(\frac{1}{n}\right)^{2/d+1}$,

$$P_n(S'(x,T_{\pi(x)}M,h)) \leq 2P(S'(x,T_{\pi(x)}M,h)) + \frac{4\,(2/d+1)\log(8n)}{n-1},$$

hence, denoting by ω_r the volume of the r-dimensional unit ball, we get

$$\begin{aligned} P_n(S(X_1,T,h)) &\leq 2Q'(\mathcal{B}(X_1,h/2)) + \frac{4\,(2/d+1)\log(8n)}{n-1} \\ &\leq \frac{2(1-\beta)\omega_d\omega_{D-d}}{K_0^D\omega_D}(k_1'h)^d(k_2'h^2)^{D-d} + \frac{4\,(2/d+1)\log(8n)}{n-1} \\ &\leq (1-\beta)C_{d,D,f_{min},\tau_{min}}h^{d+1} + \frac{4\,(2/d+1)\log(8n)}{n-1} \\ &\leq h^d\left((1-\beta)C_{d,D,\tau_{min},f_{min}}h + \frac{C_d\beta}{\kappa}\right), \end{aligned}$$

when n is large enough.

Now, if $X_1 \in M$ and $\angle\left(T_{\pi(X_1)}M, T\right) \leq Kh/\tau_{min}$, Lemma IV.18 entails that $\mathcal{B}(X_1,k_3h)\cap M \subset S(X_1,T,h)$, hence $P_n(S(X_1,T,h)) \geq P_n(\mathcal{B}(X_1,k_3h)\cap M)$. A last application of Theorem 5.1 in [BBL05] yields that, for all $h_\infty \leq h \leq h_0$, with probability larger than $1-\left(\frac{1}{n}\right)^{2/d+1}$,

$$P_n(\mathcal{B}(X_1,k_3h)\cap M) \geq \frac{1}{2}P(\mathcal{B}(X_1,k_3h)) - \frac{2\,(2/d+1)\log(8n)}{n-1}.$$

Thus we deduce that

$$\begin{aligned} P_n(S(X_1,T,h)) \geq \frac{\beta}{2}Q(\mathcal{B}(X_1,k_3h)) - \frac{2\,(2/d+1)\log(8n)}{n-1} &\geq \frac{\beta}{2}q(k_3h) - C_d\frac{\beta h^d}{\kappa} \\ &\geq h^d\left(\beta c_{d,f_{min},\tau_{min}} - C_d\frac{\beta}{\kappa}\right), \end{aligned}$$

according to Lemma B.12 (since $k_3 \leq 1$). Choosing κ large enough (depending on d, τ_{min} and f_{min}) and then n large enough leads to the result. □

B.4 Matrix Decomposition and Principal Angles

In this section we expose a standard matrix perturbation result, adapted to our framework. For real symmetric matrices, we let $\mu_i(\cdot)$ denote their i-th largest eigenvalue and $\mu_{min}(\cdot)$ the smallest one.

Theorem B.16 (Sin θ theorem [DK70], this version from Lemma 19 in [ACLZ17])**.** *Let $O \in \mathbb{R}^{D\times D}$, $B \in \mathbb{R}^{d\times d}$ be positive semi-definite symmetric matrices such that*

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

Let T_0 (resp. T) be the vector space spanned by the first d vectors of the canonical basis (resp. by the first d eigenvectors of O). Then

$$\angle(T_0,T) \leq \frac{\sqrt{2}\|E\|_{op}}{\mu_{min}(B)}.$$

B.5 Local PCA for Tangent Space Estimation and Decluttering

This section is dedicated to the proofs of Section IV.5. We begin with the case of additive noise (and no outliers), that is Proposition IV.15.

B.5.1 Proof of Proposition IV.15

Without loss of generality, the local PCA analysis will be conducted at base point X_1, the results on the whole sample then follow from a standard union bound. For convenience, we assume that $\pi_M(X_1) = 0$ and that T_0M is spanned by the d first vectors of the canonical basis of $\mathbb{R}^D$. We recall that $X_i = Y_i + Z_i$, with $Y_i \in M$ and $\|Z_i\| \leq sh$, for $s \leq 1/4$. In particular, $\|X_1\| \leq \|Z_1\| \leq sh \leq h/4$.

We adopt the following notation for the local covariance matrix based on the whole sample $\mathbb{X}_n$.

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

Note that the tangent space estimator $\texttt{TSE}(\mathbb{X}_n, h)_1$ is the space spanned by the first d eigenvectors of $\hat{\Sigma}(h)$. From now on we suppose that all the inequalities of Lemma B.14 are satisfied, defining then a global event of probability larger than $1 - 4\left(\frac{1}{n}\right)^{2/d+1}$.

We consider $h = h_0 \leq h_+$, so that Lemma B.12 and B.13 hold. We may then decompose the local covariance matrix as follows.

$$\begin{aligned} \hat{\Sigma}(h) &= \frac{1}{n-1}\sum_{i\geq 2}(X_i - m(h))(X_i - m(h))^t U(X_i,h) - \frac{N(h)}{n-1}(\bar{X}(h) - m(h))(\bar{X}(h) - m(h))^t \\ &:= \hat{\Sigma}_1 + \hat{\Sigma}_2. \end{aligned} \tag{B.17}$$

The first term may be written as

$$\begin{aligned} \hat{\Sigma}_1 &= \frac{1}{n-1}\sum_{i\geq 2}(X_i - m(h))(X_i - m(h))^t U(X_i,h) \\ &= \frac{1}{n-1}\sum_{i\geq 2}(X_i - m(h))_\top (X_i - m(h))_\top^t U(X_i,h) + R_1 \\ &= \Sigma(h) + R_1 + R_2, \end{aligned}$$

where

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

According to Lemma B.13 (with $\beta = 1$), $\mu_{min}(A(h)) \geq c_d f_{min} h^{d+2}$. On the other hand, using Proposition B.3 and Lemma B.13 we may write

$$\begin{aligned} (n-1)\|R_1\|_{\mathcal{F}}/N(h) &\leq 2 \sup_{y+z\in\mathcal{B}(X_1,h)} \|(y+z-m(h))_\top\| \|(y+z-m(h))_\perp\| \\ &\qquad + \sup_{y+z\in\mathcal{B}(X_1,h)} \|(y+z-m(h))_\perp\|^2 \\ &\leq 2 \sup_{y+z\in\mathcal{B}(X_1,h)} \|(y+z-m(h))\| \left(\|(y-m(h))_\perp\| + sh\right) \\ &\qquad + \sup_{y\in\mathcal{B}(0,2h)\cap M} \left(\|(y-m(h))_\perp\| + sh\right)^2 \\ &\leq 8h\left(\frac{4h^2}{\tau_{min}} + 2sh\right) + \left(\frac{4h^2}{\tau_{min}} + 2sh\right)^2 \\ &\leq \frac{34h^3}{\tau_{min}} + 20sh^2, \end{aligned}$$

since $h \leq h_+$ and $s \leq 1/4$. In addition, we can write

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

with $\|R_2\|_{\mathcal{F}} \leq C_d \frac{f_{max}}{f_{min}\sqrt{\kappa}} q(2h)h^2$ according to Lemma B.14 (with $\beta = 1$).

In turn, the term $\hat{\Sigma}_2$ may be decomposed as

$$\hat{\Sigma}_2 = \begin{pmatrix} R_4 & 0 \\ 0 & 0 \end{pmatrix} + R_3,$$

with

$$\begin{aligned}
\|R_4\|_{\mathcal{F}} &\leq \frac{N(h)}{n-1} \|(\bar{X}(h) - m(h))_\top\| \|(\bar{X}(h) - m(h))\| \\
&\leq \frac{2h}{n-1} \left\| \sum_{i \geq 2} (X_i - m(h))_\top U(X_i, h) \right\| \\
&\leq \frac{2C_d q(2h) h^2 f_{max}}{f_{min}\sqrt{\kappa}},
\end{aligned}$$

according to Lemma B.14. A similar bound on R_3 may be derived,

$$\begin{aligned}
\|R_3\|_{\mathcal{F}} &\leq \frac{N(h)}{n-1} \|(\bar{X}(h) - m(h))_\perp\| \|(\bar{X}(h) - m(h))\| \\
&\leq \frac{4h}{n-1} \left\| \sum_{i \geq 2} (Y_i + Z_i - m(h))_\perp U(X_i, h) \right\| \\
&\leq \frac{8hN(h)\left(2h^2/\tau_{min} + sh\right)}{n-1} \\
&\leq \frac{N(h)h^2}{n-1} \left(\frac{16h}{\tau_{min}} + 8s \right),
\end{aligned}$$

according to Proposition B.3 and Lemma B.13. If we choose $h = \left(\kappa \frac{\log n}{n-1}\right)^{1/d}$, for κ large enough (depending on d, f_{min} and f_{max}), we have

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

Now, provided that $\kappa \geq 1$, according to Lemma B.14, we may write

$$\frac{\|R_1 + R_3\|_{\mathcal{F}}}{\mu_{min}(A(h))} \leq K_{f_{max}, f_{min}, d} \left(h/\tau_{min} + s \right),$$

which, for n large enough, leads to

$$\angle(T_0 M, \hat{T}_{X_1} M) \leq \sqrt{2} K_{f_{max}, f_{min}, d} \left(h/\tau_{min} + s \right),$$

according to Proposition B.16.

B.5.2 Proof of Proposition IV.19

The proof of Proposition IV.19 follows the same path as the derivation of Proposition IV.15, with some technical difficulties due to the outliers ($\beta < 1$). We emphasize that in this framework, there is no additive noise ($\sigma = 0$). As in the previous section, the analysis will be conducted for $X_1 \in \mathbb{X}^{(k)}$, for some fixed $k \geq -1$, $k = -1$ referring to the initialization step. Results on the whole sample then follow from a standard union bound. As before, we assume that $\pi_M(X_1) = 0$ and that T_0M is spanned by the d first vectors of the canonical basis of $\mathbb{R}^D$. In what follows, denote by $\hat{t}$ the map from $\mathbb{R}^D$ to $\{0, 1\}$ such that $\hat{t}(X_i) = 1$ if and only if X_i is in $\mathbb{X}^{(k)}$.

We adopt the following notation for the local covariance matrix based on $\mathbb{X}^{(k)}$ (after $k+1$ iterations of the outlier filtering procedure).

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

Also recall that we define $N_0(h)$ and $N_1(h)$ as the number of points drawn from respectively clutter and signal in $\mathcal{B}(X_1, h) \cap M$ (based on the whole sample $\mathbb{X}_n$). At last, we suppose that all the inequalities of Lemma B.14 and Lemma B.15 are satisfied, defining then a global event of probability larger than $1 - 7\left(\frac{1}{n}\right)^{2/d+1}$.

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

$$\begin{aligned}
\hat{\Sigma}^{(k)}(h) &= \frac{1}{n-1}\sum_{i\geq 2}(X_i - m(h))(X_i - m(h))^t U(X_i, h)\hat{t}(X_i) \\
&\qquad - \frac{N^{(k)}(h)}{n-1}(\bar{X}^{(k)}(h) - m(h))(\bar{X}(h)^{(k)} - m(h))^t \\
&= \frac{1}{n-1}\sum_{i\geq 2}(X_i - m(h))(X_i - m(h))^t U(X_i, h)\hat{t}(X_i)V_i(X_i) \\
&\quad + \frac{1}{n-1}\sum_{i\geq 2}(X_i - m(h))(X_i - m(h))^t U(X_i, h)(1 - V_i)\hat{t}(X_i) \\
&\qquad\qquad - \frac{N^{(k)}(h)}{n-1}(\bar{X}(h)^{(k)} - m(h))(\bar{X}(h)^{(k)} - m(h))^t, \\
&:= \hat{\Sigma}_1^{(k)} + \hat{\Sigma}_2^{(k)} + \hat{\Sigma}_3^{(k)}.
\end{aligned} \tag{B.18}$$

The proof of Proposition IV.19 will follow by induction.

Initialization step ($k = -1$):

In this case $\mathbb{X}^{(k)} = \mathbb{X}_n$, $h = h_0$, $d(X_1, M) \leq h_0/\sqrt{2}$, and $\hat{t}$ is always equal to 1. Then the first term $\hat{\Sigma}_1^{(k)}$ of (B.18) may be written as

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

where

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

According to Lemma B.13, $\mu_{min}(A(h)) \geq c_d f_{min} \beta h^{d+2}$, and $\|R_1\|_{\mathcal{F}} \leq 34 \frac{N_1(h)h^3}{\tau_{min}(n-1)}$ according to Proposition B.3. Moreover, we can write

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

with $\|R_2\|_{\mathcal{F}} \leq C_d \frac{f_{max}}{f_{min}\sqrt{\kappa}} \beta q(2h) h^2$ according to Lemma B.14.

Term $\hat{\Sigma}_2^{(k)}$ in inequality (B.18) may be bounded by

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

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

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

with

$$\begin{aligned}
\|R_6\|_{\mathcal{F}} &\leq \frac{N(h)^{(k)}}{n-1} \|(\bar{X}(h)^{(k)} - m(h))_{\top}\| \|(\bar{X}(h)^{(k)} - m(h))\| \\
&\leq \frac{4h}{n-1} \left(\|\sum_{i \geq 2} (X_i - m(h))_{\top} U(X_i, h) V_i\| + \|\sum_{i \geq 2} (X_i - m(h))_{\top} U(X_i, h)(1 - V_i)\| \right) \\
&\leq \frac{4C_d \beta q(2h) h^2 f_{max}}{f_{min}\sqrt{\kappa}} + \frac{16h^2 N_0(h)}{n-1},
\end{aligned}$$

according to Lemma B.14. We may also write

$$\begin{aligned}
\|R_5\|_{\mathcal{F}} &\leq \frac{N(h)^{(k)}}{n-1} \|(\bar{X}(h)^{(k)} - m(h))_{\perp}\| \|(\bar{X}(h)^{(k)} - m(h))\| \\
&\leq \frac{4h}{n-1} \left(\|\sum_{i \geq 2} (X_i - m(h))_{\perp} U(X_i, h) V_i\| + \|\sum_{i \geq 2} (X_i - m(h))_{\perp} U(X_i, h)(1 - V_i)\| \right) \\
&\leq \frac{16 N_1(h) h^3}{(n-1)\tau_{min}} + \frac{16 N_0(h) h^2}{(n-1)},
\end{aligned}$$

according to Proposition B.3 and Lemma B.13. As in the additive noise case (see proof of Proposition IV.15), provided that κ is large enough (depending on d, f_{min}, and f_{max}), we have

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

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

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

according to Lemma B.14. Then, Proposition B.16 can be applied to obtain

$$\angle(\texttt{TSE}(\mathbb{X}^{(-1)}, h_0)_1, T_{\pi(X_1)} M) \leq \sqrt{2} K^{(0)}_{d, f_{min}, f_{max}, \beta} h_0 / \tau_{min}.$$

According to Lemma B.15, we may choose κ large enough (with respect to $K = \sqrt{2}K^{(0)}$, d, f_{min} and τ_{min}) and then a threshold t so that, if $X_1 \in M$, then $X_1 \in \mathbb{X}^{(0)}$, and if $d(X_1, M) \geq h_0^2/\tau_{min}$, then $X_1 \notin \mathbb{X}^{(0)}$.

Iteration step Now we assume that $k \geq 0$, and that $d(X_i, M) \geq h_k^2/\tau_{min}$ implies $\hat{t}(X_i) = 0$, with $h_k = \left(\kappa \frac{\log n}{\beta(n-1)}\right)^{\gamma_k}$, γ_k being between $1/(d+1)$ and $1/d$. Let $h_\infty \leq h \leq h_k$, and suppose that $d(X_1, M) \leq h_k/\sqrt{2}$. As in the initialization step, $\hat{\Sigma}_1^{(k)}$ may be written as

$$\begin{pmatrix} A(h) & 0 \\ 0 & 0 \end{pmatrix} + R_1 + R_2,$$

with $\mu_{min}(A(h)) \geq c_d f_{min} \beta h^{d+2}$, $\|R_1\|_{\mathcal{F}} \leq 34 \frac{N_1(h) h^3}{\tau_{min}(n-1)}$, and $\|R_2\|_{\mathcal{F}} \leq C_d \frac{f_{max}}{f_{min}\sqrt{\kappa}} \beta q(2h) h^2$.

We can decompose $\hat{\Sigma}_2$ as

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

with $\|R_4\|_{\mathcal{F}} \leq \frac{16 N_0(h) h^2}{n-1}$ and $\|R_3\| \leq \frac{128 N_0(h) h h_k^2}{(n-1)\tau_{min}}$, according to Proposition B.6, for n large enough so that $h_0^2/\tau_{min} \leq h_\infty$. Term $\hat{\Sigma}_3^{(k)}$ may also be written as

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

with

$$\begin{aligned}
\|R_6\|_{\mathcal{F}} &\leq \frac{N(h)^{(k)}}{n-1} \|(\bar{X}(h)^{(k)} - m(h))_\top\| \|(\bar{X}(h)^{(k)} - m(h))\| \\
&\leq \frac{4h}{n-1} \left(\|\sum_{i \geq 2} (X_i - m(h))_\top U(X_i, h) V_i\| + \|\sum_{i \geq 2} (X_i - m(h))_\top U(X_i, h)(1 - V_i)\hat{t}(X_i)\| \right) \\
&\leq \frac{4 C_d \beta q(2h) h^2 f_{max}}{f_{min}\sqrt{\kappa}} + \frac{16 h^2 N_0(h)}{(n-1)},
\end{aligned}$$

according to Lemma B.14. We may also write

$$\begin{aligned}
\|R_5\|_{\mathcal{F}} &\leq \frac{N(h)^{(k)}}{n-1} \|(\bar{X}(h)^{(k)} - m(h))_\perp\| \|(\bar{X}(h)^{(k)} - m(h))\| \\
&\leq \frac{4h}{n-1} \left(\|\sum_{i \geq 2} (X_i - m(h))_\perp U(X_i, h) V_i\| + \|\sum_{i \geq 2} (X_i - m(h))_\perp U(X_i, h)(1 - V_i)\hat{t}(X_i)\| \right) \\
&\leq \frac{16 N_1(h) h^3}{(n-1)\tau_{min}} + \frac{32 N_0(h) h h_k^2}{\tau_{min}(n-1)},
\end{aligned}$$

according to Proposition B.3, Proposition B.6 and Lemma B.13. As done before, we may choose κ large enough (depending on d, f_{min} and f_{max}, but not on k) such that

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

Now choose $h = h_{k+1} = \left(\kappa \frac{\log n}{\beta(n-1)}\right)^{(2\gamma_k+1)/(d+2)}$, with $\kappa \geq 1$. This choice is made to optimize residual terms of the form $h/\tau_{min} + h_k^2 N_0(h)/h$ coming from $\|R_1 + R_3 + R_5\|_{\mathcal{F}}/\mu_{min}(A(h_{k+1}))$. Then we get, according to Lemma B.14,

$$\begin{aligned} \frac{\|R_1 + R_3 + R_5\|_{\mathcal{F}}}{\mu_{min}(A(h_{k+1}))} &\leq C_d \frac{f_{max} h_{k+1}}{\tau_{min} f_{min}} + \frac{C'_d}{\beta \tau_{min} f_{min}} \left(\kappa \frac{\log n}{\beta(n-1)}\right)^{\gamma_{k+1}+2\gamma_k-(2\gamma_k+1)+1} \\ &\leq K_{d,f_{min},f_{max},\beta} \frac{h_{k+1}}{\tau_{min}}, \end{aligned} \tag{B.19}$$

where again, $K_{d,f_{min},f_{max},\beta}$ does not depend on k. At last, we may apply Proposition B.16 to get

$$\begin{aligned} \angle(\texttt{TSE}(\mathbb{X}^{(k)}, h_{k+1})_1, T_{\pi(X_1)}M) &\leq \sqrt{2} K_{d,f_{min},f_{max},\beta} h_{k+1}/\tau_{min} \\ &\leq \sqrt{2}\left(K_{d,f_{min},f_{max},\beta} \vee K^{(0)}_{d,f_{min},f_{max},\beta}\right) h_{k+1}/\tau_{min} \\ &:= C_{d,\beta,f_{max},f_{min}} h_{k+1}/\tau_{min}. \end{aligned}$$

Then, according to Lemma B.15, we may choose κ large enough (not depending on k) and t (not depending on k either) so that if $X_1 \in M$, then $X_1 \in \mathbb{X}^{(k+1)}$, and if $d(X_1, M) \geq h_k^2/\tau_{min}$, then $X_1 \notin \mathbb{X}^{(k+1)}$. Proposition IV.19 then follows from a straightforward union bound on the sample $\{X_1, \ldots, X_n\}$.

B.5.3 Proof of Proposition IV.22

In this case, we have $d(X_j, M) \leq h_\infty^2/\tau_{min}$, for every X_j in $\mathbb{X}^{(\hat{k})}$. The proof of Proposition IV.22 follows from the same calculation as in the proof of Proposition IV.19, replacing h_k^2/τ_{min} by its upper bound h_∞^2/τ_{min} and taking $h_{k+1} = h_\infty$ in the iteration step.

B.6 Proof of the Main Reconstruction Results

We now prove main results Theorem IV.7 (additive noise model), and Theorems IV.8 and IV.9 (clutter noise model).

B.6.1 Additive Noise Model

Proof of Corollary IV.16. Let $Q \in \mathcal{P}^2_{\tau_{min},\sigma}(f_{min}, f_{max})$. Write $\varepsilon = c_{d,f_{min},f_{max}}(h \vee \tau_{min}\sigma/h)$ for $c_{d,f_{min},f_{max}}$ large enough, an consider the event A defined by

$$\begin{aligned} A = &\left\{ \max_{X_j \in \mathbb{X}_n} \angle(T_{\pi_M(X_j)}M, \hat{T}_j(h)) \leq C_{d,f_{min},f_{max}} \left(\frac{h}{\tau_{min}} + \frac{\sigma}{h}\right) \right\} \cap \left\{ \sup_{x \in M} d(x, \mathbb{X}_n) \leq \sigma \right\} \\ &\cap \left\{ \sup_{X_j \in \mathbb{X}_n} d(X_j, M) \leq C_{d,f_{min}} \left(\frac{\log n}{n}\right)^{1/d} \right\}. \end{aligned}$$

Then from Proposition IV.15 and Lemma B.11, $\mathbb{P}_Q(A) \geq 1 - 5\left(\frac{1}{n}\right)^{2/d}$, and from the definition of ε and the construction of $\mathbb{Y}_n$, for n large enough,

$$\begin{aligned} A &\subset \left\{ \max_{X_j \in \mathbb{X}_n} \angle(T_{\pi_M(X_j)}M, \hat{T}_j(h)) \leq \frac{\varepsilon}{2280\tau_{min}} \right\} \cap \left\{ \sup_{x \in M} d(x, \mathbb{X}_n) \leq \varepsilon \right\} \\ &\quad \cap \left\{ \sup_{X_j \in \mathbb{X}_n} d(X_j, M) \leq \frac{\varepsilon^2}{1140\tau_{min}} \right\} \\ &\subset \left\{ \max_{X_j \in \mathbb{Y}_n} \angle(T_{\pi_M(X_j)}M, \hat{T}_j(h)) \leq \frac{\varepsilon}{2280\tau_{min}} \right\} \cap \left\{ \sup_{x \in M} d(x, \mathbb{Y}_n) \leq 2\varepsilon \right\} \\ &\quad \cap \{\mathbb{Y}_n \text{ is } \varepsilon\text{-sparse}\} \cap \left\{ \sup_{X_j \in \mathbb{Y}_n} d(X_j, M) \leq \frac{\varepsilon^2}{1140\tau_{min}} \right\}, \end{aligned}$$

which yields the result.

□

Proof of Theorem IV.7. Following the above notation, we observe that on the event A, Theorem IV.14 holds for $\varepsilon = c_{d,f_{min},f_{max}}(h \vee \tau_{min}\sigma/h)$, $\theta = \varepsilon/(1140\tau_{min})$ (where we used that $\theta \leq 2\sin\theta$) and $\eta = \varepsilon^2/(1140\tau_{min})$ with high probability, so that the first part of Theorem IV.7 is proved. Furthermore, for n large enough,

$$\begin{aligned} \mathbb{E}_{Q^n}\left[d_H\left(M, \hat{M}_{\texttt{TDC}}\right)\right] &= \mathbb{E}_Q\left[d_H\left(M, \hat{M}_{\texttt{TDC}}\right)\mathbb{1}_A\right] + \mathbb{E}_Q\left[d_H\left(M, \hat{M}_{\texttt{TDC}}\right)\mathbb{1}_{A^c}\right] \\ &\leq C_d\frac{\varepsilon^2}{\tau_{min}} + (1 - \mathbb{P}_Q(A))\,(\mathrm{diam}(M) + \sigma) \\ &\leq C'_{d,f_{min},f_{max},\tau_{min}}\varepsilon^2, \end{aligned}$$

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

B.6.2 Clutter Noise Model

Proof of Corollary IV.20. Let $P \in \mathcal{P}^2_{\tau_{min},\beta}(f_{min}, f_{max})$. For n large enough, write $\varepsilon = c_{d,f_{min}f_{max}}h_{k_\delta}$ for $c_{d,f_{min}f_{max}}$ large enough, and consider the event

$$\begin{aligned} A^\delta = &\left\{ \max_{X_j \in \mathbb{X}^{(k_\delta)}} \angle(T_{\pi_M(X_j)}M, \hat{T}^\delta_j) \leq C_{d,f_{min},f_{max}}\frac{h_{k_\delta}}{\tau_{min}} \right\} \cap \left\{ \sup_{x \in M} d(x, \mathbb{X}^{(k_\delta)}) \leq \frac{h^2_{k_\delta}}{\tau_{min}} \right\} \\ &\cap \left\{ \sup_{X_j \in \mathbb{X}^{(k_\delta)}} d(X_j, M) \leq C_{d,f_{min}}\left(\frac{\log n}{n}\right)^{1/d} \right\}. \end{aligned}$$

From Proposition IV.19 and Lemma B.11, $\mathbb{P}_P(A^\delta) \geq 1 - 8\left(\frac{1}{n}\right)^{2/d}$ and from the definition of ε and the construction of $\mathbb{Y}^\delta_n$, for n large enough,

$$
\begin{aligned}
A^\delta &\subset \left\{ \max_{X_j \in \mathbb{X}^{(k_\delta)}} \angle(T_{\pi_M(X_j)}M, \hat{T}_j^\delta) \leq \frac{\varepsilon}{2280\tau_{min}} \right\} \cap \left\{ \sup_{x \in M} d(x, \mathbb{X}^{(k_\delta)}) \leq \varepsilon \right\} \\
&\quad \cap \left\{ \sup_{X_j \in \mathbb{X}^{(k_\delta)}} d(X_j, M) \leq \frac{\varepsilon^2}{1140\tau_{min}} \right\} \\
&\subset \left\{ \max_{X_j \in \mathbb{Y}_n^\delta} \angle(T_{\pi_M(X_j)}M, \hat{T}_j^\delta) \leq \frac{\varepsilon}{2280\tau_{min}} \right\} \cap \left\{ \sup_{x \in M} d(x, \mathbb{Y}_n^\delta) \leq 2\varepsilon \right\} \\
&\quad \cap \{\mathbb{Y}_n \text{ is } \varepsilon\text{-sparse}\} \cap \left\{ \sup_{X_j \in \mathbb{Y}_n^\delta} d(X_j, M) \leq \frac{\varepsilon^2}{1140\tau_{min}} \right\},
\end{aligned}
$$

which yields the result.

□

Proof of Theorem IV.8. Following the above notation, we observe that on the event A^δ, Theorem IV.14 holds for $\varepsilon = c_{d,f_{min}f_{max}} h_{k_\delta}$, $\theta = \varepsilon/(1140\tau_{min})$ and $\eta = \varepsilon^2/(1140\tau_{min})$, so that the first part of Theorem IV.8 is proved. As a consequence, for n large enough,

$$
\begin{aligned}
\mathbb{E}_{P^n}\left[d_H\left(M, \hat{M}_{\texttt{TDC}\delta}\right)\right] &= \mathbb{E}_P\left[d_H\left(M, \hat{M}_{\texttt{TDC}\delta}\right) \mathbb{1}_{A^\delta}\right] + \mathbb{E}_P\left[d_H\left(M, \hat{M}_{\texttt{TDC}\delta}\right) \mathbb{1}_{\left(A^\delta\right)^c}\right] \\
&\leq C_d \frac{\varepsilon^2}{\tau_{min}} + \left(1 - \mathbb{P}_P(A^\delta)\right) \times 2K_0 \\
&\leq C'_{d,f_{min},f_{max},\tau_{min}} \varepsilon^2,
\end{aligned}
$$

where for the second line we used the fact that $M \cup \hat{M}_{\texttt{TDC}\delta} \subset \mathcal{B}_0$, a ball of radius $K_0 = K_0(d, f_{min}, \tau_{min})$. □

Finally, Theorem IV.9 is obtained similarly using Proposition IV.22.

Chapter V

Approximation and Geometry of the Reach

Abstract

As illustrated in Chapter IV, various problems in manifold estimation make use of the *reach*, denoted by τ_M, which is a measure of the regularity of the submanifold. This chapter is the first investigation into the problem of how to estimate the reach. First, we study the geometry of the reach through an approximation perspective. We derive new geometric results on the reach for submanifolds without boundary. An estimator $\hat{\tau}$ of τ_M is proposed in a framework where tangent spaces are known, and bounds assessing its efficiency are derived. In the case of i.i.d. random point cloud $\mathbb{X}_n$, $\hat{\tau}(\mathbb{X}_n)$ is showed to achieve uniform expected loss bounds over a $\mathcal{C}^3$-like model. Finally, we obtain upper and lower bounds on the minimax rate for estimating the reach.

Content

V.1 Introduction

Manifold estimation has become an increasingly important problem in statistics and machine learning. There is now a large literature on methods and theory for estimating manifolds. See, for example, [KZ15, GPPVW12a, FMN16, BG14, NSW08, BNS06, GK06]. Estimating a manifold, or functionals of a manifold, requires regularity conditions. In nonparametric function estimation, regularity conditions often take the form of smoothness constraints. In manifold estimation problems, as illustrated in Chapter IV, a common assumption is that the reach τ_M of the manifold M is bounded away from zero.

First introduced by Federer [Fed59], the reach τ_M of a set $M \subset \mathbb{R}^D$ is the largest number such that any point at distance less than τ_M from M has a unique nearest point on M. If a set has its reach greater than $\tau_{min} > 0$, then one can roll freely a ball of radius τ_{min} around it [CFPL12]. The reach is affected by two factors: the curvature of the manifold and the width of the narrowest bottleneck-like structure of M, which quantifies how close M is from being self-intersecting.

Positive reach is the minimal regularity assumption on sets in geometric measure theory and integral geometry [Fed69, Thä08]. Sets with positive reach exhibit a structure that is close to being differential — the so-called tangent and normal cones. The value of the reach itself quantifies the degree of regularity of a set, with larger values associated to more regular sets. The positive reach assumption is routinely imposed in the statistical analysis of geometric structures in order to ensure good statistical properties [CFPL12] and to derive theoretical guarantees. For example, in manifold reconstruction, the reach helps formalize minimax rates [GPPVW12a, KZ15]. The optimal manifold estimators of Chapter IV implicitly use reach as a scale parameter in their construction. In homology inference [NSW08, BRSW13], the reach drives the minimal sample size required to consistently estimate topological invariants. The reach is used in [CFRC07] as a regularity parameter in the estimation of the Minkowski boundary lengths and surface areas. The reach has been explicitly used as a regularity parameter in geometric inference, such as in volume estimation [APR16] and manifold clustering [ACLZ17]. The reach is also used as a scale parameter in dimension reduction techniques such as vector diffusions maps [SW12]. Problems in computational geometry such as manifold reconstruction also rely on assumptions on the reach [BG14].

In this chapter we study the problem of estimating reach. To do so, we first provide new geometric results on the reach. We also give the first bounds on the minimax rate for estimating reach.

There are very few papers on this problem. When the embedding dimension is 3, the estimation of the local feature size (a localized version of the reach) was tackled in a deterministic way in [DS06]. To some extent, the estimation of the medial axis (the set of points that have strictly more than one nearest point on M) and its generalizations [CLPL14, ABE09] can be viewed as an indirect way to estimate the reach. A test procedure designed to validate whether data actually comes from a smooth manifold satisfying a condition on the reach was developed in [FMN16]. The authors derived a consistent test procedure, but the results do not permit any inference bound on the reach.

Outline

In Section V.2 we provide some differential geometric background and define the statistical problem at hand. New geometric properties of the reach are derived in Section V.3, and their consequences for its inference follow Section V.4 in a setting where tangent spaces are known. We study minimax rates in Section V.5. An extension to a model where tangent

spaces are unknown is discussed Section V.6, and we conclude with some open questions in Section V.7.

V.2 Framework

V.2.1 Notation

In what follows, $D \geq 2$ and $\mathbb{R}^D$ is endowed with the Euclidean scalar product $\langle \cdot, \cdot \rangle$ and the associated norm $\|\cdot\|$. The associated closed ball of radius r and center x is denoted by $\mathcal{B}(x, r)$. We will consider compact connected submanifolds M of $\mathbb{R}^D$ of fixed dimension $1 \leq d < D$ and without boundary [dC92]. For every point p in M, the tangent space of M at p is denoted by T_pM. It is the d-dimensional vector subspace of $\mathbb{R}^D$ composed of the directions locally spanned by M at p. Besides the Euclidean structure given by $\mathbb{R}^D \supset M$, a submanifold is endowed with an intrinsic metric structure induced by the ambient Euclidean one, called the geodesic distance. Given a smooth path $c : [a, b] \to M$, the length of c is defined as $Length(c) = \int_a^b \|c'(t)\| \, dt$. One can show [dC92] that there exists a path γ of minimal length joining p and q. Such an arc is called geodesic, and the geodesic distance between p and q is given by $d_M(p, q) = \inf_{c(0)=a, c(1)=b} Length(c)$. We let $\mathcal{B}_M(p, s)$ denote the closed geodesic ball of center $p \in M$ and of radius s. A geodesic γ such that $\|\gamma'(t)\| = 1$ for all t is called arc-length parametrized. Unless stated otherwise, a geodesic will always be considered in its arc-length version. For all $p \in M$ and all unit vectors $v \in T_pM$, we denote by $\gamma_{p,v}$ the unique arc-length parametrized geodesic of M such that $\gamma_{p,v}(0) = p$ and $\gamma'_{p,v}(0) = v$. The exponential map is defined as $\exp_p(vt) = \gamma_{p,v}(t)$. Note that from the compactness of M, $\exp_p : T_pM \to M$ is defined globally on T_pM. For any two nonzero vectors $u, v \in \mathbb{R}^D$, we let $\angle(u, v) = d_{\mathcal{S}^{D-1}}(\frac{u}{\|u\|}, \frac{v}{\|v\|})$ be the angle between u and v.

V.2.2 Reach

First introduced by Federer [Fed59], the reach regularity parameter is defined as follows. Given a closed subset $A \subset \mathbb{R}^D$, the medial axis $Med(A)$ of A is the subset of $\mathbb{R}^D$ consisting of the points that have at least two nearest neighbors on A. Namely, denoting by $d(z, A) = \inf_{p \in A} \|p - z\|$ the distance function to A,

$$Med(A) = \left\{ z \in \mathbb{R}^D | \exists p \neq q \in A, \|p - z\| = \|q - z\| = d(z, A) \right\}. \tag{V.1}$$

The reach of A is then defined as the minimal distance from A to $Med(A)$.

Definition V.2. *The reach of a closed subset $A \subset \mathbb{R}^D$ is defined as*

$$\tau_A = \inf_{p \in A} d\left(p, Med(A)\right) = \inf_{z \in Med(A)} d\left(z, A\right). \tag{V.3}$$

Some authors refer to τ_A^{-1} as the *condition number* [NSW08, SW12]. From the definition of the medial axis in (V.1), the projection $\pi_A(x) = \arg\min_{p \in A} \|p - x\|$ onto A is well defined outside $Med(A)$. The reach is the largest distance $\rho \geq 0$ such that π_A is well defined on the ρ-offset $\{x \in \mathbb{R}^D | d(x, A) \leq \rho\}$. Hence, the reach condition can be seen as a generalization of convexity, since a set $A \subset \mathbb{R}^D$ is convex if and only if $\tau_A = \infty$.

In the case of submanifolds, one can reformulate the definition of the reach in the following manner.

Theorem V.4 (Theorem 4.18 in [Fed59]). *For all submanifolds $M \subset \mathbb{R}^D$,*

$$\tau_M = \inf_{q \neq p \in M} \frac{\|q-p\|^2}{2d(q-p, T_pM)}. \tag{V.5}$$

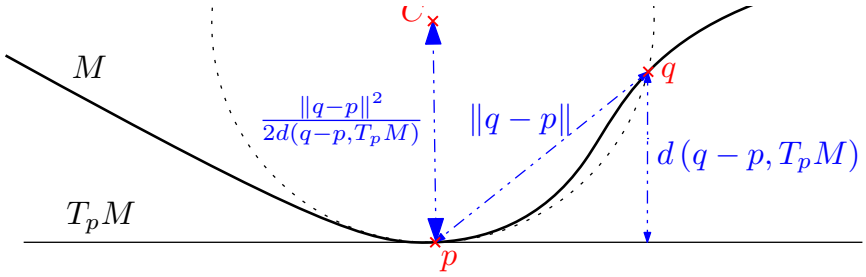


Figure V.1 – Geometric interpretation of quantities involved in (V.5).

This formulation has the advantage of involving only points on M and its tangents spaces, while (V.3) uses the distance to the medial axis $Med(M)$, which is a global quantity. The formula (V.5) will be the starting point of the estimator proposed in this chapter (see Section V.4).

The ratio appearing in (V.5) can be interpreted geometrically, as suggested in Figure V.1. This ratio is the radius of an ambient ball, tangent to M at p and passing through q. Hence, at a differential level, the reach gives a lower bound on the radii of curvature of M. Equivalently, τ_M^{-1} is an upper bound on the curvature of M. The following result was already reproduce in Proposition III.20, though we state it here for sake of completeness.

Proposition V.6 (Proposition 6.1 in [NSW08]). *Let $M \subset \mathbb{R}^D$ be a submanifold, and $\gamma_{p,v}$ an arc-length parametrized geodesic of M. Then for all t,*

$$\left\|\gamma''_{p,v}(t)\right\| \leq 1/\tau_M.$$

In analogy with function spaces, the class $\left\{M \subset \mathbb{R}^D | \tau_M \geq \tau_{min} > 0\right\}$ can be interpreted as the Hölder space $\mathcal{C}^2(1/\tau_{min})$. In addition, as illustrated in Figure V.2, the condition $\tau_M \geq \tau_{min} > 0$ also prevents bottleneck structures where M is nearly self-intersecting. This idea will be made rigorous in Section V.3.

V.2.3 Statistical Model and Loss

Let us now describe the regularity assumptions we will use throughout. To avoid arbitrarily irregular shapes, we consider submanifolds M with their reach lower bounded by $\tau_{min} > 0$. Since the parameter of interest τ_M is a $\mathcal{C}^2$-like quantity, it is natural — and actually

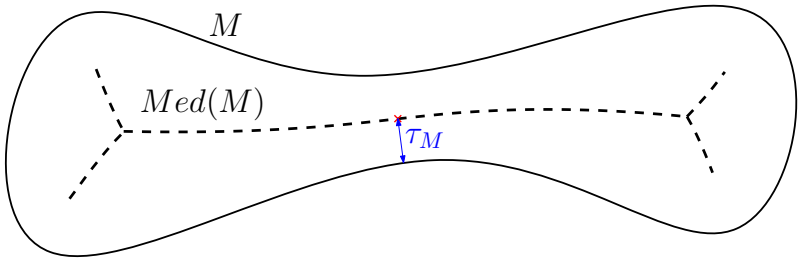


Figure V.2 – A narrow bottleneck structure yields a small reach τ_M.

necessary from Proposition V.13 — to require an extra degree of smoothness. For example, by imposing an upper bound on the third order derivatives of geodesics.

Definition V.7. *We let $\mathcal{C}^{(3)}_{\tau_{min},L}$ denote the set of compact connected d-dimensional submanifolds $M \subset \mathbb{R}^D$ without boundary such that $\tau_M \geq \tau_{min}$, and for which every arc-length parametrized geodesic $\gamma_{p,v}$ is $\mathcal{C}^3$ and satisfies*

$$\left\|\gamma'''_{p,v}(0)\right\| \leq L.$$

Note that since the third order condition $\left\|\gamma'''_{p,v}(0)\right\| \leq L$ needs to hold for all (p, v), we have in particular that $\left\|\gamma'''_{p,v}(t)\right\| \leq L$ for all $t \in \mathbb{R}$. To our knowledge, such a $\mathcal{C}^3$ quantitative assumption on the geodesics has not been considered in the computational geometry literature. Here, we chose the notation with "(3)" in parentheses to emphasize the fact that the $\mathcal{C}^3$ assumption we do deals with geodesic trajectories, and not with parametrizations of the manifold. Additionally, we will consider another $\mathcal{C}^3$ assumption in Chapter VI with this notation.

Any submanifold $M \subset \mathbb{R}^D$ of dimension d inherits a natural measure vol_M from the d-dimensional Hausdorff measure $\mathcal{H}^d$ on $\mathbb{R}^D$ [Fed69, p. 171]. We will consider distributions P that have densities with respect to vol_M that are bounded away from zero.

Definition V.8. *We let $\mathcal{P}^{(3)}_{\tau_{\min},L}(f_{\min})$ denote the set of distributions P having support $M \in \mathcal{C}^{(3)}_{\tau_{min},L}$ and with a Hausdorff density $f = \frac{dP}{dvol_M}$ satisfying $\inf_{x\in M} f(x) \geq f_{\min} > 0$ on M.*

Notice that, with the notation of Chapter IV, we have

$$\mathcal{P}^{(3)}_{\tau_{\min},L=\infty}(f_{\min}) = \mathcal{P}^{2}_{\tau_{\min}}(f_{\min}, f_{max} = \infty).$$

That is, setting $L = \infty$ boils down to consider a $\mathcal{C}^2$ model.

In order to focus on the geometric aspects of the reach, we will first consider the case where tangent spaces are observed at all the sample points. We let $\mathbb{G}^{d,D}$ denote the Grassmanian of dimension d of $\mathbb{R}^D$, that is the set of all d-dimensional vector subspaces of $\mathbb{R}^D$.

Definition V.9. *For any distribution $P \in \mathcal{P}^{(3)}_{\tau_{\min},L}(f_{\min})$ with support M we associate the distribution $\tilde{P}$ of the random variable $(X, T_X M)$ on $\mathbb{R}^D \times \mathbb{G}^{d,D}$, where X has distribution P. We let $\tilde{\mathcal{P}}^{(3)}_{\tau_{\min},L}(f_{\min})$ denote the set of all such distributions.*

Formally, one can write $\tilde{P}(dx\, dT) = \delta_{T_x M}(dT)P(dx)$, where $\delta_{\cdot}$ denotes the Dirac measure. An i.i.d. n-sample of P is of the form $(X_1, T_1), \ldots, (X_n, T_n) \in \mathbb{R}^D \times \mathbb{G}^{d,D}$, where $X_1, \ldots, X_n$ is an i.i.d. n-sample of P and $T_i = T_{X_i} M$ with $M = Supp(P)$. For a distribution P with support M and associated distribution $\tilde{P}$ on $\mathbb{R}^D \times \mathbb{G}^{d,D}$, we will write $\tau_{\tilde{P}} = \tau_P = \tau_M$, with a slight abuse of notation.

Note that the model does not explicitly impose an upper bound on τ_M. Such an upper bound would be redundant, since the lower bound on f_{min} does impose such an upper bound, as stated in Proposition III.26, that we reproduce here for sake of completeness.

Proposition V.10 (Proposition III.26)**.** *Let $M \subset \mathbb{R}^D$ be a connected closed d-dimensional manifold, and let P be a probability distribution with support M. Assume that P has a density f with respect to the Hausdorff measure on M such that $\inf_{x\in M} f(x) \geq f_{min} > 0$. Then,*

$$\tau_M^d \leq \frac{C_d}{f_{min}},$$

for some constant $C_d > 0$ depending only on d.

To simplify the statements and the proofs, we focus on a loss involving the condition number. Namely, we measure the error with the loss

$$\ell(\tau, \tau') = \left| \frac{1}{\tau} - \frac{1}{\tau'} \right|^p, \quad p \geq 1. \tag{V.11}$$

In other words, we will consider the estimation of the condition number τ_M^{-1} instead of the reach τ_M.

Remark V.12. *For a distribution $P \in \mathcal{P}^{(3)}_{\tau_{\min},L}(f_{\min})$, Proposition V.10 asserts that $\tau_{min} \leq \tau_P \leq \tau_{max} := (C_d/f_{min})^{1/d}$. Therefore, in an inference set-up, we can always restrict to estimators $\hat{\tau}$ within the bounds $\tau_{min} \leq \hat{\tau} \leq \tau_{max}$. Consequently,*

$$\frac{1}{\tau_{max}^{2p}} |\tau_P - \hat{\tau}|^p \leq \left| \frac{1}{\tau_P} - \frac{1}{\hat{\tau}} \right|^p \leq \frac{1}{\tau_{min}^{2p}} |\tau_P - \hat{\tau}|^p ,$$

so that the estimation of the reach τ_P is equivalent to the estimation of the condition number τ_P^{-1}, up to constants.

With the statistical framework developed above, we can now see explicitly why the third order condition $\|\gamma'''\| \leq L < \infty$ is necessary. Indeed, the next result demonstrates how relaxing this constraint — *i.e.* setting $L = \infty$ — renders the problem of reach estimation intractable. Below, σ_d stands for the volume of the d-dimensional unit sphere $\mathcal{S}^d$.

Proposition V.13. *For all $\tau_{min} > 0$, provided that $f_{min} \geq \frac{1}{2^{d+1}\tau_{min}^d \sigma_d}$, for all $n \geq 1$,*

$$\inf_{\hat{\tau}_n} \sup_{\tilde{P} \in \tilde{\mathcal{P}}^{(3)}_{\tau_{min}, L=\infty}(f_{min})} \mathbb{E}_{\tilde{P}^n} \left| \frac{1}{\tau_{\tilde{P}}} - \frac{1}{\hat{\tau}_n} \right|^p \geq \frac{c_p}{\tau_{min}^p} > 0,$$

where the infimum is taken over the estimators $\hat{\tau}_n = \hat{\tau}_n(X_1, T_1, \ldots, X_n, T_n)$.

Thus, one cannot expect to derive uniformly good approximation bounds solely under the condition $\tau_M \geq \tau_{min}$. This result is natural, since the problem at stake is to estimate a differential quantity of order two. Therefore, some notion of uniform $\mathcal{C}^3$ regularity is needed.

V.3 Geometry of the Reach

In this section, we give a precise geometric description of how the reach arises. In particular, below we will show that the reach is determined either by a bottleneck structure or an area of high curvature (Theorem V.17). These two cases are referred to as *global* reach and *local* reach, respectively.

Consider the formulation (V.3) of the reach as the infimum of the distance between M and its medial axis $Med(M)$. By definition of the medial axis (V.1), if the infimum is attained it corresponds to a point z_0 in $Med(M)$ at distance τ_M from M, which we call an *axis point*. Since z_0 belongs to the medial axis of M, it has at least two nearest neighbors q_1, q_2 on M, which we call a *reach attaining pair* (see Figure V.3(b)). By definition, q_1 and q_2 belong to $\mathcal{B}(z_0, \tau_M)$ and cannot be farther than $2\tau_M$ from each other. We say that (q_1, q_2) is a *bottleneck* of M in the extremal case $\|q_2 - q_1\| = 2\tau_M$ of antipodal points of $\mathcal{B}(z_0, \tau_M)$ (see Figure V.3(a)). Note that the ball $\mathcal{B}(z_0, \tau_M)$ meets M only on its boundary $\partial\mathcal{B}(z_0, \tau_M)$.

Definition V.14. *Let $M \subset \mathbb{R}^D$ be a submanifold with reach $\tau_M > 0$.*

- *A pair of points* (q_1, q_2) *in* M *is called reach attaining if there exists* $z_0 \in Med(M)$ *such that* $q_1, q_2 \in \mathcal{B}(z_0, \tau_M)$. *We call* z_0 *the axis point of* (q_1, q_2), *and* $\|q_1 - q_2\| \in (0, 2\tau_M]$ *its size.*

- *A reach attaining pair* $(q_1, q_2) \in M^2$ *is said to be a bottleneck of* M *if its size is* $2\tau_M$, *that is* $\|q_1 - q_2\| = 2\tau_M$.

As stated in the following Lemma V.15, if a reach attaining pair is not a bottleneck — that is $\|q_1 - q_2\| < 2\tau_M$ —, then M contains an arc of a circle of radius τ_M. In this sense, this "semi-local" case — when $\|q_1 - q_2\|$ can be arbitrarily small — is not generic. Though, we do not exclude this case in the analysis.

Lemma V.15. *Let* $M \subset \mathbb{R}^D$ *be a compact submanifold with reach* $\tau_M > 0$. *Assume that* M *has a reach attaining pair* $(q_1, q_2) \in M^2$ *with size* $\|q_1 - q_2\| < 2\tau_M$. *Let* $z_0 \in Med(M)$ *be their associated axis point, and write* $c_{z_0}(q_1, q_2)$ *for the arc of the circle with center* z_0 *and endpoints as* q_1 *and* q_2.

Then $c_{z_0}(q_1, q_2) \subset M$, *and this arc (which has constant curvature* $1/\tau_M$*) is the geodesic joining* q_1 *and* q_2.

In particular, in this "semi-local" situation, since τ_M^{-1} is the norm of the second derivative of a geodesic of M (the exhibited arc of the circle of radius τ_M), the reach can be viewed as arising from directional curvature.

Now consider the case where the infimum (V.3) is not attained. In this case, the following Lemma V.16 asserts that τ_M is created by curvature.

Lemma V.16. *Let* $M \subset \mathbb{R}^D$ *be a compact submanifold with reach* $\tau_M > 0$. *Assume that for all* $z \in Med(M)$, $d(z, M) > \tau_M$. *Then there exists* $q_0 \in M$ *and a geodesic* γ_0 *such that* $\gamma_0(0) = q_0$ *and* $\|\gamma_0''(0)\| = \frac{1}{\tau_M}$.

To summarize, there are three distinct geometric instances in which the reach may be realized:

- M has a bottleneck: by definition, τ_M originates from a structure having scale $2\tau_M$ (see Figure V.3(a)).

- M has a reach attaining pair but no bottleneck: then M contains an arc of a circle of radius τ_M (Lemma V.15), so that M actually contains a zone with radius of curvature τ_M (see Figure V.3(b)).

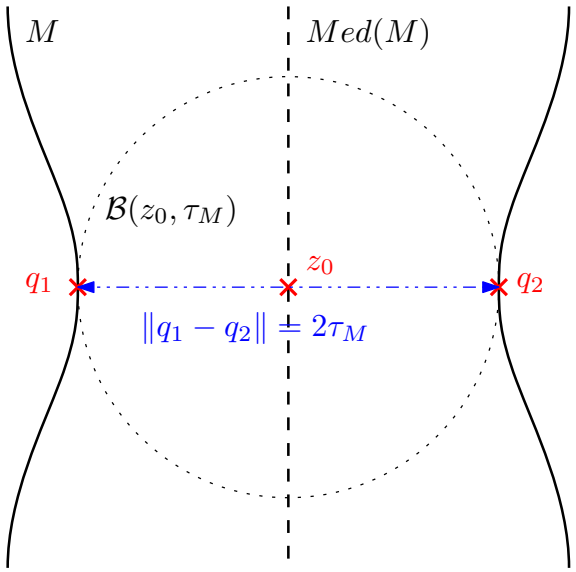
- M does not have a reach attaining pair: then τ_M originates from curvature (Lemma V.16), also yielding a point with radius of curvature τ_M. (see Figure V.3(c)).

From now on, we will treat the first case separately from the other two. We are now in a position to state the main result of this section. It is a straightforward consequence of Lemma V.15 and Lemma V.16.

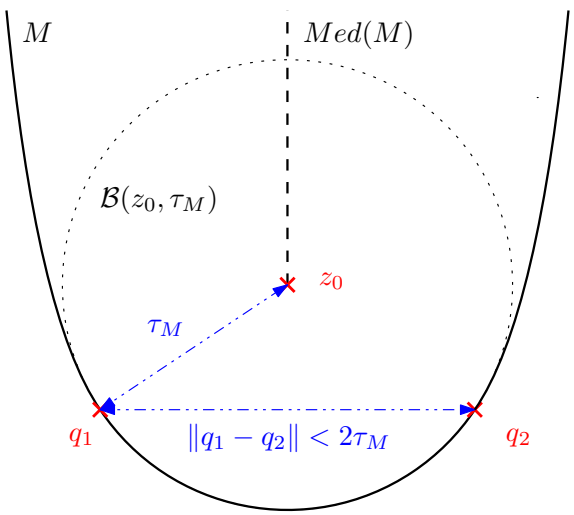
Theorem V.17. *Let* $M \subset \mathbb{R}^D$ *be a compact submanifold with reach* $\tau_M > 0$. *At least one of the following two assertions holds.*

- (Global case) M *has a bottleneck* $(q_1, q_2) \in M^2$, *that is, there exists* $z_0 \in Med(M)$ *such that* $q_1, q_2 \in \partial\mathcal{B}(z_0, \tau_M)$ *and* $\|q_1 - q_2\| = 2\tau_M$.

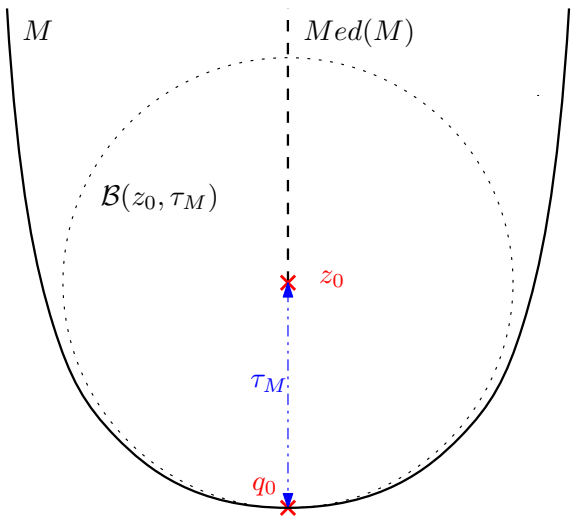
- (Local case) *There exists* $q_0 \in M$ *and an arc-length parametrized geodesic* γ_0 *such that* $\gamma_0(0) = q_0$ *and* $\|\gamma_0''(0)\| = \frac{1}{\tau_M}$.



(a) A bottleneck.



(b) A reach attaining pair but no bottleneck.



(c) No reach attaining pair.

Figure V.3 – The different ways for the reach to be attained.

Theorem V.17 provides a description of the reach as arising from global and local geometric structures that, to the best of our knowledge, is new. Such a distinction is especially important in our problem. Indeed, the global and local cases may yield different approximation properties and require different statistical analyses. However, since one does not know a priori whether the reach arises from a global or a local structure, an estimator of τ_M should be able to handle both cases simultaneously.

V.4 Reach Estimator and its Analysis

In this section, we propose an estimator $\hat{\tau}(\cdot)$ for the reach and demonstrate its properties and rate of consistency under the loss (V.11). We rely on the formulation of the reach given in (V.5) (see also Figure V.1), and define $\hat{\tau}$ as a plugin estimator as follows. Given a point cloud $\mathcal{X} = \{x_1, \ldots, x_n\} \subset M$, we let

$$\hat{\tau}(\mathcal{X}) = \inf_{x \neq y \in \mathcal{X}} \frac{\|y - x\|^2}{2d(y - x, T_x M)}. \tag{V.18}$$

In particular, we have $\hat{\tau}(M) = \tau_M$. Since the infimum (V.18) is taken over a set $\mathcal{X}$ smaller than M, $\hat{\tau}(\mathcal{X})$ always overestimates τ_M. In fact, $\hat{\tau}(\mathcal{X})$ is decreasing in the number of distinct points in $\mathcal{X}$, a useful property that we formalize in the following result, whose proof is immediate.

Corollary V.19. *Let M be a submanifold with reach τ_M and $\mathcal{Y} \subset \mathcal{X} \subset M$ be two nested subsets. Then $\hat{\tau}(\mathcal{Y}) \geq \hat{\tau}(\mathcal{X}) \geq \tau_M$.*

We now derive the rate of consistency of $\hat{\tau}$. We analyze the global case (Section V.4.1) and the local case (Section V.4.2) separately. In both cases, we first determine the performance of the estimator in a deterministic framework, and then derive an expected loss bounds when $\hat{\tau}$ is applied to a random sample.

V.4.1 Global Case

Consider the global case, that is, M has a bottleneck structure (Theorem V.17). Then the infimum (V.5) is achieved at a bottleneck pair $(q_1, q_2) \in M^2$. When $\mathcal{X}$ contains points that are close to q_1 and q_2, one may expect that the infimum over the sample points should also be close to (V.5): that is, that $\hat{\tau}(\mathcal{X})$ should be close to τ_M.

Proposition V.20. *Let $M \subset \mathbb{R}^D$ be a submanifold with reach $\tau_M > 0$ that has a bottleneck $(q_1, q_2) \in M^2$ (Definition V.14), and $\mathcal{X} \subset M$. If there exist $x, y \in \mathcal{X}$ with $\|q_1 - x\| < \tau_M$ and $\|q_2 - y\| < \tau_M$, then*

$$0 \leq \frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathcal{X})} \leq \frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\{x, y\})} \leq \frac{9}{2\tau_M^2} \max\left\{d_M(q_1, x), d_M(q_2, y)\right\}.$$

The error made by $\hat{\tau}(\mathcal{X})$ decreases linearly in the maximum of the distances to the critical points q_1 and q_2. In other words, the radius of the tangent sphere in Figure V.1 grows at most linearly in t when we perturb by $t < \tau_M$ its basis point $p = q_1$ and the point $q = q_2$ it passes through.

Based on the deterministic bound of Proposition V.20, we can now give an upper bound on the expected loss under the model $\mathcal{P}^{(3)}_{\tau_{\min}, L}(f_{\min})$. We recall that, here and in what follows, $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ is an i.i.d. sample with common distribution P.

Proposition V.21. *Let $P \in \mathcal{P}^{(3)}_{\tau_{\min},L}(f_{\min})$ and $M = Supp(P)$. Assume that M has a bottleneck $(q_1, q_2) \in M^2$ (see Definition V.14). Then,*

$$\mathbb{E}_{P^n}\left[\left|\frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathbb{X}_n)}\right|^p\right] \leq C_{p,d,\tau_M,f_{\min}} n^{-\frac{p}{d}},$$

where $C_{p,d,\tau_{\min},f_{\min}}$ depends only on p,d,τ_M and $f_{\min}$, and is a decreasing function of τ_M.

Let us emphasize the fact that although not explicit in the notation, $\hat{\tau}(\mathbb{X}_n)$ depends on the tangent spaces of M at the points of $\mathbb{X}_n$. Proposition V.21 follows straightforwardly from Proposition V.20 combined with the fact that with high probability, the balls centered at the bottleneck points q_1 and q_2 with radii $\mathcal{O}(n^{-1/d})$ both contain a sample point of $\mathbb{X}_n$.

V.4.2 Local Case

Consider now the local case, that is, there exists $q_0 \in M$ and $v_0 \in T_{q_0}M$ such that the geodesic $\gamma_0 = \gamma_{q_0,v_0}$ has second derivative $\|\gamma_0''(0)\| = 1/\tau_M$ (Theorem V.17). Estimating τ_M boils down to estimating the curvature of M at q_0 in the direction v_0.

We first relate directional curvature to the increment $\frac{\|y-x\|^2}{2d(y-x,T_xM)}$ involved in the estimator $\hat{\tau}$ (V.18). Indeed, since the latter quantity is the radius of a sphere tangent at x and passing through y (Figure V.1), it approximates the radius of curvature in the direction $y - x$ when x and y are close. For $x, y \in M$, we let $\gamma_{x\to y}$ denote the arc-length parametrized geodesic joining x and y, with the convention $\gamma_{x\to y}(0) = x$.

Lemma V.22. *Let $M \in \mathcal{C}^{(3)}_{\tau_{\min},L}$ with reach τ_M and $\mathcal{X} \subset M$ be a subset. Let $x, y \in \mathcal{X}$ with $d_M(x, y) < \pi\tau_M$. Then,*

$$0 \leq \frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathcal{X})} \leq \frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\{x,y\})} \leq \frac{1}{\tau_M} - \left\|\gamma''_{x\to y}(0)\right\| + \frac{2}{3}Ld_M(x,y).$$

Let us now state how directional curvatures are stable with respect to perturbations of the base point and the direction. We let κ_p denote the maximal directional curvature of M at $p \in M$, that is,

$$\kappa_p = \sup_{v\in\mathcal{B}_{T_pM}(0,1)} \left\|\gamma''_{p,v}(0)\right\|.$$

Lemma V.23. *Let $M \in \mathcal{C}^{(3)}_{\tau_{\min},L}$ with reach τ_M and $q_0, x, y \in M$ be such that $x, y \in \mathcal{B}_M\left(q_0, \frac{\pi\tau_M}{2}\right)$. Let γ_0 be a geodesic such that $\gamma_0(0) = q_0$ and $\|\gamma_0''(0)\| = \kappa_{q_0}$. Write*

$$\theta_x := \angle(\gamma_0'(0), \gamma'_{q_0\to x}(0)), \quad \theta_y := \angle(\gamma_0'(0), \gamma'_{q_0\to y}(0)),$$

and suppose that $|\theta_x - \theta_y| \geq \frac{\pi}{2}$. Then,

$$\begin{aligned}&\left\|\gamma''_{x\to y}(0)\right\| \\ &\geq \kappa_{q_0} - \frac{1}{\sqrt{2}-1}\left(\kappa_x - \kappa_{q_0} + \sqrt{2}(3\kappa_{q_0} + \kappa_x)\sin^2(|\theta_x - \theta_y|) + \sqrt{2}Ld_M(q_0, x)\right).\end{aligned}$$

In particular, geodesics in a neighborhood of q_0 with directions close to v_0 have curvature close to $\frac{1}{\tau_M}$. A point cloud $\mathcal{X}$ sampled densely enough in M would contain points in this neighborhood. Hence combining Lemma V.22 and Lemma V.23 yields the following deterministic bound in the local case.

Proposition V.24. *Let $M \in \mathcal{C}^{(3)}_{\tau_{\min},L}$ be such that there exist $q_0 \in M$ and a geodesic γ_0 such that $\gamma_0(0) = q_0$ and $\|\gamma_0''(0)\| = \frac{1}{\tau_M}$. Let $\mathcal{X} \subset M$ and $x, y \in \mathcal{X}$ be such that $x, y \in \mathcal{B}_M\left(q_0, \frac{\pi\tau_M}{2}\right)$. Let*

$$\theta_x := \angle(\gamma_0'(0), \gamma'_{q_0 \to x}(0)), \quad \theta_y := \angle(\gamma_0'(0), \gamma'_{q_0 \to y}(0)),$$

and suppose that $|\theta_x - \theta_y| \geq \frac{\pi}{2}$. Then,

$$\begin{aligned} 0 \leq \frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathcal{X})} &\leq \frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\{x,y\})} \\ &\leq \frac{4\sqrt{2}\sin^2(|\theta_x - \theta_y|)}{(\sqrt{2}-1)\tau_M} + L\left(\frac{2}{3}d_M(x,y) + \frac{\sqrt{2}}{\sqrt{2}-1}d_M(q_0,x)\right). \end{aligned}$$

In other words, since the reach boils down to directional curvature in the local case, $\hat{\tau}$ performs well if it is given as input a pair of points x, y which are close to the point q_0 realizing the reach, and almost aligned with the direction of interest v_0.

Similarly to the analysis of the global case, the deterministic bound of Proposition V.24 yields a bound on the risk of $\hat{\tau}(\mathbb{X}_n)$ when $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ is random.

Proposition V.25. *Let $P \in \mathcal{P}^{(3)}_{\tau_{\min},L}(f_{\min})$ and $M = Supp(P)$. Suppose there exists $q_0 \in M$ and a geodesic γ_0 with $\gamma_0(0) = q_0$ and $\|\gamma_0''(0)\| = \frac{1}{\tau_M}$. Then,*

$$\mathbb{E}_{P^n}\left[\left|\frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathbb{X}_n)}\right|^p\right] \leq C_{\tau_{\min},d,L,f_{\min},p} n^{-\frac{2p}{3d-1}},$$

where $C_{\tau_{\min},d,L,f_{\min},p}$ depends only on $\tau_{\min}, d, L, f_{\min}$ and p.

This statement follows from Proposition V.24 together with the estimate of the probability of two points being drawn in a neighborhood of q_0 and subject to an alignment constraint.

Proposition V.21 and V.25 yield a convergence rate of $\hat{\tau}(\mathbb{X}_n)$ which is slower in the local case than in the global case. Recall that from Theorem V.17, the reach pertains to the size of a bottleneck structure in the global case, and to maximum directional curvature in the local case. To estimate the size of a bottleneck, observing two points close to each point in the bottleneck gives a good approximation. However, for approximating maximal directional curvature, observing two points close to the curvature attaining point is not enough, but they should also be aligned with the highly curved direction. Hence, estimating the reach may be more difficult in the local case, and the difference in the convergence rates of Proposition V.21 and V.25 matches this intuition.

V.5 Minimax Estimates

In this section we derive bounds on the minimax risk R_n of the estimation of the reach over the class $\tilde{\mathcal{P}}^{(3)}_{\tau_{\min},L}(f_{\min})$, that is

$$R_n = \inf_{\hat{\tau}_n} \sup_{\tilde{P} \in \tilde{\mathcal{P}}^{(3)}_{\tau_{\min},L}(f_{min})} \mathbb{E}_{\tilde{P}^n}\left|\frac{1}{\tau_{\tilde{P}}} - \frac{1}{\hat{\tau}_n}\right|^p, \tag{V.26}$$

where the infimum ranges over all estimators $\hat{\tau}_n((X_1, T_{X_1}), \ldots, (X_n, T_{X_n}))$ based on an i.i.d. sample of size n with the knowledge of the tangent spaces at sample points.

The rate of consistency of the plugin estimator $\hat{\tau}(\mathbb{X}_n)$ studied in the previous section leads to an upper bound on R_n, which we state here for completeness.

Theorem V.27. *For all $n \geq 1$,*

$$R_n \leq C_{\tau_{\min},d,L,f_{\min},p} n^{-\frac{2p}{3d-1}},$$

for some constant $C_{\tau_{\min},d,L,f_{\min},p}$ depending only on $\tau_{\min}, d, L, f_{\min}$ and p.

We now focus on deriving a lower bound on the minimax risk R_n. The method relies on an application of Le Cam's Lemma [Yu97]. In what follows, let

$$TV\left(Q, Q'\right) = \frac{1}{2}\int |dQ - dQ'|$$

denote the total variation distance between Q and Q', where dQ, dQ' denote the respective densities of Q, Q' with respect to any dominating measure. Since $|x-z|^p + |z-y|^p \geq 2^{1-p}|x-y|^p$, the following version of Le Cam's lemma results from Lemma 1 in [Yu97] and $(1 - TV(Q^n, Q'^n)) \geq (1 - TV(Q, Q'))^n$.

Lemma V.28 (Le Cam's Lemma)**.** *Let $\tilde{P}, \tilde{P}' \in \tilde{\mathcal{P}}^{(3)}_{\tau_{\min},L}(f_{\min})$ with respective supports M and M'. Then for all $n \geq 1$,*

$$R_n \geq \frac{1}{2^p}\left|\frac{1}{\tau_M} - \frac{1}{\tau_{M'}}\right|^p \left(1 - TV(\tilde{P}, \tilde{P}')\right)^n.$$

Lemma V.28 implies that in order to derive a lower bound on R_n one needs to consider distributions (hypotheses) in the model that are stochastically close to each other — i.e. with small total variation distance — but for which the associated reaches are as different as possible. A lower bound on the minimax risk over $\tilde{\mathcal{P}}^{(3)}_{\tau_{\min},L}(f_{\min})$ requires the hypotheses to belong to the class. Luckily, in our problem it will be enough to construct hypotheses from the simpler class $\mathcal{P}^{(3)}_{\tau_{\min},L}(f_{\min})$. Indeed, we have the following isometry result between $\mathcal{P}^{(3)}_{\tau_{\min},L}(f_{\min})$ and $\tilde{\mathcal{P}}^{(3)}_{\tau_{\min},L}(f_{\min})$ for the total variation distance.

Lemma V.29. *In accordance with the notation of Definition V.9, let $P, P' \in \mathcal{P}^{(3)}_{\tau_{\min},L}(f_{\min})$ be distributions on $\mathbb{R}^D$ with associated distributions $\tilde{P}, \tilde{P}' \in \tilde{\mathcal{P}}^{(3)}_{\tau_{\min},L}(f_{\min})$ on $\mathbb{R}^D \times \mathbb{G}^{d,D}$. Then,*

$$TV\left(\tilde{P}, \tilde{P}'\right) = TV\left(P, P'\right).$$

In order to construct hypotheses in $\mathcal{P}^{(3)}_{\tau_{\min},L}(f_{\min})$ we take advantage of the fact that the class $\mathcal{C}^{(3)}_{\tau_{min},L}$ has good stability properties, which we now describe. Here, since submanifolds do not have natural parametrizations, the notion of perturbation can be well formalized using diffeomorphisms of the ambient space $\mathbb{R}^D \supset M$. Given a smooth map $\Phi : \mathbb{R}^D \to \mathbb{R}^D$, we denote by $d^i_x\Phi$ its differential of order i at x. Given a tensor field A between Euclidean spaces, let $\|A\|_{op} = \sup_x \|A_x\|_{op}$, where $\|A_x\|_{op}$ is the operator norm induced by the Euclidean norm. The next result states, informally, that the reach and geodesics third derivatives of a submanifold that is perturbed by a diffeomorphism that is $\mathcal{C}^3$-close to the identity map do not change much.

Proposition V.30. *Let $M \in \mathcal{C}^{(3)}_{\tau_{min}L}$ be fixed, and let $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ be a global $\mathcal{C}^3$-diffeomorphism. If $\|I_D - d\Phi\|_{op}$, $\left\|d^2\Phi\right\|_{op}$ and $\left\|d^3\Phi\right\|_{op}$ are small enough, then $M' = \Phi(M) \in \mathcal{C}^{(3)}_{\frac{\tau_{min}}{2},2L}$.*

Now we construct the two hypotheses P, P' as follows (see Figure V.4). Take M to be a d-dimensional sphere and P to be the uniform distribution on it. Let $M' = \Phi(M)$, where Φ is a bump-like diffeomorphism having the curvature of M' to be different of that of M in some small neighborhood. Finally, let P' be the uniform distribution on M'.

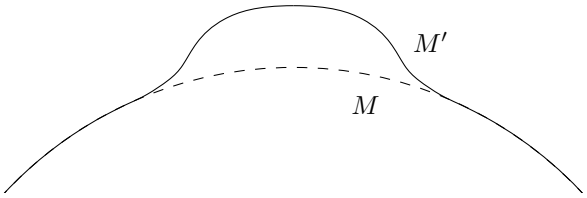


Figure V.4 – Hypotheses of Proposition V.31

Proposition V.31. *Assume that* $L \geq \frac{1}{2\tau_{min}^2}$ *and* $f_{min} \geq \frac{1}{2^{d+1}\tau_{min}^d \sigma_d}$. *Then for* $\ell > 0$ *small enough, there exist* $P, P' \in \mathcal{P}^{(3)}_{\tau_{\min},L}(f_{min})$ *with respective supports* M *and* M' *such that*

$$\left|\frac{1}{\tau_M} - \frac{1}{\tau_{M'}}\right| \geq c_d \frac{\ell}{\tau_{min}^2} \quad \text{and} \quad TV\left(P, P'\right) \leq 12\left(\frac{\ell}{2\tau_{min}}\right)^d.$$

Hence, applying Lemma V.28 with the hypotheses $\tilde{P}, \tilde{P}'$ associated to P, P' of Proposition V.31, and taking $12\left(\ell/2\tau_{min}\right)^d = 1/n$, together with Lemma V.29, yields the following lower bound.

Proposition V.32. *Assume that* $L \geq \frac{1}{2\tau_{min}^2}$ *and* $f_{min} \geq \frac{1}{2^{d+1}\tau_{min}^d \sigma_d}$. *Then for* n *large enough,*

$$R_n \geq \frac{c_{d,p}}{\tau_{min}^p} n^{-p/d},$$

where $c_{d,p}$ *depends only on* d *and* p.

Here, the assumptions on the parameters L and f_{min} are necessary for the model to be rich enough. Roughly speaking, they ensure at least that a sphere of radius $2\tau_{min}$ belongs to the model. From Proposition V.32, the plugin estimation $\hat{\tau}(\mathbb{X}_n)$ provably achieves the optimal rate in the global case (Theorem V.21) up to numerical constants. In the local case (Theorem V.25) the rate obtained presents a gap, yielding a gap in the overall rate.

V.6 Towards Unknown Tangent Spaces

So far, in our analysis we have used the key assumption that both the point cloud and the tangent spaces were jointly observed. We now focus on the more realistic framework where only points are observed. We once again rely on the formulation of the reach given in Theorem V.5 and consider a new plug-in estimator in which the true tangent spaces are replaced by estimated ones. Namely, given a point cloud $\mathcal{X} \subset \mathbb{R}^D$ and a family $T = \{T_x\}_{x \in \mathcal{X}}$ of linear subspaces of $\mathbb{R}^D$ indexed by $\mathcal{X}$, the estimator is defined as

$$\hat{\tau}(\mathcal{X}, T) = \inf_{x \neq y \in \mathcal{X}} \frac{\|y - x\|^2}{2d(y - x, T_x)}. \tag{V.33}$$

In particular, $\hat{\tau}(\mathcal{X}) = \hat{\tau}(\mathcal{X}, T_{\mathcal{X}}M)$, where $T_{\mathcal{X}}M = \{T_xM\}_{x \in \mathcal{X}}$. Adding uncertainty on tangent spaces in (V.33) does not change drastically the estimator, as the formula is stable with respect to T. In what follows, the distance between two vector subspaces $U, V \in \mathbb{G}^{d,D}$ is measured with the sine of their principal angle $\|\pi_U - \pi_V\|_{op}$.

Proposition V.34. *Let* $\mathcal{X} \subset \mathbb{R}^D$ *and* $T = \{T_x\}_{x \in \mathcal{X}}$, $T' = \{T'_x\}_{x \in \mathcal{X}}$ *be two families of linear subspaces of* $\mathbb{R}^D$ *indexed by* $\mathcal{X}$. *Assume* $\mathcal{X}$ *to be* δ*-sparse,* T *and* T' *to be* θ*-close, in the sense that*

$$\inf_{x \neq y \in \mathcal{X}} \|y - x\| \geq \delta \quad \text{and} \quad \sup_{x \in \mathcal{X}} \|T_x - T'_x\|_{op} \leq \sin\theta.$$

Then,

$$\left| \frac{1}{\hat{\tau}(\mathcal{X}, T)} - \frac{1}{\hat{\tau}(\mathcal{X}, T')} \right| \leq \frac{2 \sin \theta}{\delta}.$$

In other words, the map $T \mapsto \hat{\tau}(\mathcal{X}, T)^{-1}$ is smooth, provided that the basis point cloud $\mathcal{X}$ contains no zone of accumulation at a too small scale $\delta > 0$. As a consequence, under the assumptions of Proposition V.34, the bounds on $\left|\hat{\tau}(\mathcal{X})^{-1} - {\tau_M}^{-1}\right|$ of Proposition V.20 and Proposition V.24 still hold with an extra error term $2 \sin \theta / \delta$ if we replace $\hat{\tau}(\mathcal{X})$ by $\hat{\tau}(\mathcal{X}, T)$.

For an i.i.d. point cloud $\mathbb{X}_n$ asymptotic rates of tangent space estimation derived in $\mathcal{C}^3$-like models can be found in [CC16, SW12], yielding bounds on $\sin \theta$. In that case, the typical scale of minimum interpoint distance is $\delta \asymp n^{-2/d}$, as stated in the asymptotic result Theorem 2.1 in [KMT92] for the flat case of $\mathbb{R}^d$. However, the typical covering scale of M used in the global case (Theorem V.21) is $\varepsilon \asymp (1/n)^{1/d}$. It appears that we can sparsify the point cloud $\mathbb{X}_n$ — that is, removing accumulation points — while preserving the covering property at scale $\varepsilon = 2\delta \asymp (\log n / n)^{1/d}$. This can be performed using the *farthest point sampling algorithm* (see Section IV.3.3). Such a sparsification pre-processing allows to lessen the possible instability of $\hat{\tau}(\mathbb{X}_n, \cdot)^{-1}$. Though, whether the alignment property used in the local case (Theorem V.25) is preserved under sparsification remains to be investigated.

V.7 Conclusion and Open Questions

In this chapter, we gave new insights on the geometry of the reach. Inference results were derived in both deterministic and random frameworks. For i.i.d. samples, non-asymptotic minimax upper and lower bounds were derived under assumptions on the third order derivative of geodesic trajectories. Let us conclude with some open questions.

- The minimax upper and lower bounds given in Theorem V.27 and Theorem V.32 do not match. They are yet to be sharpened.

- In practice, since large reach ensures regularity, one may be interested with having a lower bound on the reach τ_M. Giving the limiting distribution of the statistic $\hat{\tau}(\mathbb{X}_n)$ would allow to derive asymptotic confidence intervals for τ_M.

- Other regularity parameters such as local feature size [BG14] and λ-reach [CL05] could be relevant to estimate, as they are used as tuning parameters in computational geometry techniques.

Appendix C

Proofs for Chapter V

Content

C.1 Some Technical Results on the Model

This section garners geometric lemmas on embedded manifolds in the Euclidean space that are related to the reach, and that will be used several times in the proofs. For most of them, the following results were already stated in Chapter III.

Proposition C.1. *Let $M \subset \mathbb{R}^D$ be a submanifold with reach $\tau_M > 0$.*

(i) For all $p \in M$, we let II_p^M denote the second fundamental form of M at p. Then for all unit $v \in T_pM$, $\left\|II_p^M(v,v)\right\| \leq \frac{1}{\tau_M}$.

(ii) The injectivity radius of M is at least $\pi\tau_M$.

(iii) The sectional curvatures κ of M satisfy $-\frac{2}{\tau_M^2} \leq \kappa \leq \frac{1}{\tau_M^2}$.

(iv) For all $p \in M$, the map $\exp_p : \mathring{\mathcal{B}}_{T_pM}(0, \pi\tau_M) \to \mathring{\mathcal{B}}_M(0, \pi\tau_M)$ is a diffeomorphism. Moreover, for all $\|v\| < \frac{\pi\tau_M}{2\sqrt{2}}$ and $w \in T_pM$,

$$\left(1 - \frac{\|v\|^2}{6\tau_M^2}\right)\|w\| \leq \left\|d_v \exp_p \cdot w\right\| \leq \left(1 + \frac{\|v\|^2}{\tau_M^2}\right)\|w\|$$

(v) For all $p \in M$, $r \leq \frac{\pi\tau_M}{2\sqrt{2}}$, and a Borel set $A \subset \mathcal{B}_{T_qM}(0,r) \subset T_qM$,

$$\left(1 - \frac{r^2}{6\tau_M^2}\right)^d \mathcal{H}^d(A) \leq \mathcal{H}^d(\exp_q(A)) \leq \left(1 + \frac{r^2}{\tau_M^2}\right)^d \mathcal{H}^d(A).$$

(vi) Let $q \in M$, γ be a geodesic at q, and denote by P_t the parallel transport operator along γ. Then for all $t < \pi\tau_M$ and for all $v \in T_qM$,

$$\angle(P_t(v), v) \leq \frac{t}{\tau_M}.$$

Proof of Proposition C.1. For (i),(ii),(iii),(iv), and (v), see Proposition III.22. All that remain to be showed is (vi). For this, assume without loss of generality that $\|v\| = 1$. Let $g : [0,t] \to \mathcal{S}^{d-1}$ be defined by $g(s) = P_s(v)$. Let $u \in \mathbb{R}^D$ be a unit vector and denoting by $\bar{\nabla}$ the ambient derivative. We may write

$$\langle g'(s), u\rangle = \left\langle \bar{\nabla}_{\gamma'(s)} P_s(w), u \right\rangle = \left\langle II^M_{\gamma(s)}(\gamma'(s), P_s(w)), u \right\rangle.$$

Hence $\|g'(s)\| \leq \frac{1}{\tau_M}$ for all $s \in [0,t]$. Since g is a curve on $\mathcal{S}^{d-1}$, this implies

$$\angle(P_t(v), v) = d_{\mathcal{S}^{d-1}}(\gamma(t), \gamma(0)) \leq \int_0^t \|g'(s)\|\, ds \leq \frac{t}{\tau_M}.$$

□

C.2 Geometry of the Reach

For $M \subset \mathbb{R}^D$, $a \in M$, and $v \in \mathbb{R}^D$ a non-zero vector, we define the *local directional reach* by

$$reach(M, a, v) = \inf\left\{ d(x, M) | x \in \overline{Med(M)} \text{ with } x = a + tv \text{ for some } t \geq 0 \right\},$$

with the convention $reach(M, a, v) = \infty$ if $\overline{Med(M)} \cap \{a + tv | t \geq 0\} = \emptyset$.

Lemma C.2. *(i) For $x \notin Med(M) \cup M$, writing $a = \pi_M(x)$, we have $reach(M, a, x - a) > 0$, and for all $b \in M$,*

$$\langle x - a, a - b\rangle \geq -\frac{\|a - b\|^2 \|x - a\|}{2reach(M, a, x - a)}.$$

(ii) Let $0 < r < q < \infty$ be fixed. Let $x, y \notin Med(M) \cup M$ be such that $d(x, M), d(y, M) \leq r$ and

$$reach\left(M, \pi_M(x), x - \pi_M(x)\right) \geq q \ , \ reach\left(M, \pi_M(y), y - \pi_M(y)\right) \geq q.$$

Then,

$$\|\pi_M(x) - \pi_M(y)\| \leq \frac{q}{q - r} \|x - y\|.$$

Proof of Lemma C.2. (i) The proof follows that of Theorem 4.8 (7) in [Fed59]. Let $v = \frac{x-a}{\|x-a\|}$ and $S = \{t | \pi_M(a + tv) = a\}$. As $\|x - a\| > 0$ belongs to S, $\sup S > 0$ and from [Fed59, Theorem 4.8 (6)] we get

$$\sup S \geq reach(M, a, v).$$

Moreover, for $0 < t \in S$,

$$\|a + tv - b\| \geq d(a + tv, M) = t.$$

Developing and rearranging the square of previous inequality yields

$$\begin{aligned} \|a-b\|^2 + 2t\langle v, a-b\rangle + t^2 &\geq t^2, \\ 2t\langle v, a-b\rangle &\geq -\|a-b\|^2, \\ \langle x-a, a-b\rangle &\geq -\frac{\|a-b\|^2 \|x-a\|}{2t}. \end{aligned}$$

(ii) The proof follows that of Theorem 4.8 (8) in [Fed59]. Writing $a = \pi_M(x)$ and $b = \pi_M(y)$, the previous point yields,

$$\langle x-a, a-b\rangle \geq -\frac{\|a-b\|^2 r}{2q} \text{ and } \langle y-b, b-a\rangle \geq \frac{\|a-b\|^2 r}{2q}.$$

As a consequence,

$$\begin{aligned} \|x-y\|\|a-b\| &\geq \langle x-y, a-b\rangle \\ &= \langle (x-a)+(a-b)+(b-y), a-b\rangle \\ &\geq \|a-b\|^2 \left(1 - \frac{r}{q}\right), \end{aligned}$$

hence the result.

□

Lemma C.3. *Let $M \subset \mathbb{R}^D$ be a submanifold with reach $\tau_M > 0$ having a reach attaining pair $(q_1, q_2) \in M^2$ such that $\|q_1 - q_2\| < 2\tau_M$. Write $z_0 \in Med(M)$ for the associated axis point. Then there exists a sequence of curves $\{\gamma_n\}_{n\in\mathbb{N}}$ of M joining q_1 and q_2 with*

$$\lim_n Length(\gamma_n) = \tau_M \angle(q_1 - z_0, q_2 - z_0).$$

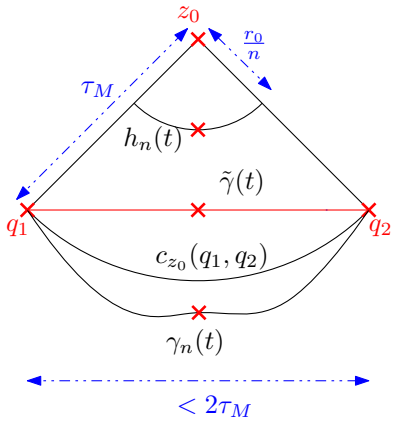


Figure C.1 – Layout of Lemma C.3.

Proof of Lemma C.3. Without loss of generality, assume that z_0 coincides with the origin. Let $c_{z_0}(q_1, q_2)$ be the circle arc of center z_0 with endpoints q_1 and q_2, and let $\gamma : [-t_0, t_0] \to c_{z_0}(q_1, q_2)$ be its arc length parametrization with $\gamma(-t_0) = q_1$ and $\gamma(t_0) = q_2$. Let

$\theta := \angle(q_1 - z_0, q_2 - z_0)$. Since $\|q_1 - z_0\| = \|q_2 - z_0\| = \tau_M$, we have $t_0 = \frac{1}{2}\tau_M\theta$. For all $t \in [-t_0, t_0]$, let $r_t := \sqrt{\tau_M^2 - \frac{\|q_1-q_2\|^2}{4}} \csc\left(\frac{t}{\tau_M}\right)$, and let $\tilde{\gamma} : [-t_0, t_0] \to \mathbb{R}^D$ be $\tilde{\gamma}(t) = \frac{r_t}{\tau_M}\gamma(t)$. Let us show that for all $r \in (0, r_0]$ and $t \in [-t_0, t_0]$, following holds:

$$\mathring{\mathcal{B}}\left(\frac{r}{\tau_M}\gamma(t), r\right) \subset \mathring{\mathcal{B}}\left(\tilde{\gamma}(t), r_t\right) \subset \mathring{\mathcal{B}}\left(q_1, \tau_M\right) \cup \mathring{\mathcal{B}}\left(q_2, \tau_M\right), \tag{C.4}$$

The left-hand side inclusion of (C.4) being trivial, we turn to the second inclusion. First, note that by definition,

$$\tilde{\gamma}(t) = \left(\frac{1}{2} - \frac{\tan\left(\frac{t}{\tau_M}\right)}{2\tan\left(\frac{t_0}{\tau_M}\right)}\right) q_1 + \left(\frac{1}{2} + \frac{\tan\left(\frac{t}{\tau_M}\right)}{2\tan\left(\frac{t_0}{\tau_M}\right)}\right) q_2$$

for all $t \in [-t_0, t_0]$. Hence,

$$\tilde{\gamma}(t) - \tilde{\gamma}(0) = \frac{\tan\left(\frac{t}{\tau_M}\right)}{2\tan\left(\frac{t_0}{\tau_M}\right)}(q_2 - q_1), \tag{C.5}$$

and from $\tan\left(\frac{t_0}{\tau_M}\right) = \frac{\|q_1-q_2\|}{2r_0}$, we get $\|\tilde{\gamma}(t) - \tilde{\gamma}(0)\| = r_0 \tan\left(\frac{t}{\tau_M}\right)$. Now suppose $x \in \mathring{\mathcal{B}}\left(\tilde{\gamma}(t), r_t\right)$, then

$$\|x - \tilde{\gamma}(t)\|^2 < r_t^2. \tag{C.6}$$

Then,

$$\|x - \tilde{\gamma}(t)\|^2 = \|x - \tilde{\gamma}(0)\|^2 - 2\left\langle x - \tilde{\gamma}(0), \tilde{\gamma}(t) - \tilde{\gamma}(0)\right\rangle + \|\tilde{\gamma}(t) - \tilde{\gamma}(0)\|^2,$$

and $r_t^2 = r_0^2 + r_0^2 \tan^2\left(\frac{t}{\tau_M}\right) = r_0^2 + \|\tilde{\gamma}(t) - \tilde{\gamma}(0)\|^2$, hence applying these and (C.5) to (C.6) implies

$$\|x - \tilde{\gamma}(0)\|^2 - \frac{\tan\left(\frac{t}{\tau_M}\right)}{\tan\left(\frac{t_0}{\tau_M}\right)}\left\langle x - \tilde{\gamma}(0), q_2 - q_1\right\rangle < r_0^2. \tag{C.7}$$

Now applying $\tilde{\gamma}(-t_0) = q_1$ to (C.5) gives $q_1 - \tilde{\gamma}(0) = -\frac{1}{2}(q_2 - q_1)$, so

$$\begin{aligned}\|x - q_1\|^2 &= \|x - \tilde{\gamma}(0)\|^2 + 2\left\langle x - \tilde{\gamma}(0), q_1 - \tilde{\gamma}(0)\right\rangle + \|q_1 - \tilde{\gamma}(0)\|^2 \\ &= \|x - \tilde{\gamma}(0)\|^2 - \left\langle x - \tilde{\gamma}(0), q_2 - q_1\right\rangle + \frac{1}{4}\|q_1 - q_2\|^2.\end{aligned}$$

Similarly,

$$\|x - q_2\|^2 = \|x - \tilde{\gamma}(0)\|^2 + \left\langle x - \tilde{\gamma}(0), q_2 - q_1\right\rangle + \frac{1}{4}\|q_1 - q_2\|^2,$$

and hence

$$\begin{aligned}&\min\left\{\|x - q_1\|^2, \|x - q_2\|^2\right\} \\ &= \|x - \tilde{\gamma}(0)\|^2 - \left|\left\langle x - \tilde{\gamma}(0), q_2 - q_1\right\rangle\right| + \frac{1}{4}\|q_1 - q_2\|^2.\end{aligned} \tag{C.8}$$

Since $\left|\tan\left(\frac{t_0}{\tau_M}\right)\right| \le \left|\tan\left(\frac{t}{\tau_M}\right)\right|$, applying (C.7) to (C.8) gives

$$\begin{aligned}&\min\left\{\|x - q_1\|^2, \|x - q_2\|^2\right\} \\ &\le \|x - \tilde{\gamma}(0)\|^2 - \frac{\tan\left(\frac{t}{\tau_M}\right)}{\tan\left(\frac{t_0}{\tau_M}\right)}\left\langle x - \tilde{\gamma}(0), q_2 - q_1\right\rangle + \frac{1}{4}\|q_1 - q_2\|^2 \\ &< r_0^2 + \frac{1}{4}\|q_1 - q_2\|^2 = \tau_M^2,\end{aligned}$$

which asserts the second inclusion in (C.4).

Now, by definition of the reach in (V.3), $\left(\mathring{\mathcal{B}}(q_1, \tau_M) \cup \mathring{\mathcal{B}}(q_2, \tau_M)\right) \cap Med(M) = \emptyset$, hence (C.4) implies

$$\mathring{\mathcal{B}}\left(\frac{r}{\tau_M}\gamma(t), r\right) \cap Med(M) = \emptyset.$$

For all $n \in \mathbb{N}$, let us now define $h_n, \gamma_n : [-t_0, t_0] \to M$ by (See Figure C.1),

$$h_n(t) = \frac{r_0}{n\tau_M}\gamma(t) \quad \text{and} \quad \gamma_n(t) = \pi_M(h_n(t)).$$

Then for any fixed $n \in \mathbb{N}$ and $t_1, t_2 \in [-t_0, t_0]$ such that $|t_1 - t_2| < \tau_M$, from $\mathring{\mathcal{B}}\left(h_n(t_i), \frac{r_0}{n}\right) \cap Med(M) = \emptyset$, we get

$$\begin{aligned} reach\left(M, \gamma_n(t_i), h_n(t_i) - \gamma_n(t_i)\right) &\geq d\left(h_n(t_i), M\right) + \frac{r_0}{n} \\ &\geq d(h_n(t_1), M) \wedge d(h_n(t_2), M) + \frac{r_0}{n}, \end{aligned}$$

and since $d(h_n(t_i), M) \leq d(h_n(t_1), M) \vee d(h_n(t_2), M)$, Lemma C.2 (ii) yields

$$\begin{aligned} &\|\gamma_n(t_1) - \gamma_n(t_2)\| = \|\pi_M(h_n(t_1)) - \pi_M(h_n(t_2))\| \\ &\leq \frac{\left(d\left(h_n(t_1), M\right) \wedge d\left(h_n(t_2), M\right) + \frac{r_0}{n}\right)\|h_n(t_1) - h_n(t_2)\|}{d\left(h_n(t_1), M\right) \wedge d\left(h_n(t_2), M\right) + \frac{r_0}{n} - d\left(h_n(t_1), M\right) \vee d\left(h_n(t_2), M\right)} \\ &= \frac{d\left(h_n(t_1), M\right) \wedge d\left(h_n(t_2), M\right) + \frac{r_0}{n}}{\frac{r_0}{n} - |d(h_n(t_1), M) - d(h_n(t_2), M)|}\|h_n(t_1) - h_n(t_2)\|. \end{aligned}$$

Noticing furthermore that

$$|d(h_n(t_1), M) - d(h_n(t_2), M)| \leq \|h_n(t_1) - h_n(t_2)\| \leq \frac{r_0}{n\tau_M}|t_1 - t_2|,$$

and

$$d(h_n(t_i), M) \leq d(z_0, M) + \|h_n(t_i) - z_0\| \leq \tau_M + \frac{r_0}{n},$$

we get

$$\begin{aligned} \|\gamma_n(t_1) - \gamma_n(t_2)\| &\leq \frac{\tau_M + 2\frac{r_0}{n}}{\frac{r_0}{n} - \frac{r_0}{n\tau_M}|t_1 - t_2|}\frac{r_0}{n\tau_M}|t_1 - t_2| \\ &= \frac{\tau_M + 2\frac{r_0}{n}}{\tau_M - |t_1 - t_2|}|t_1 - t_2|. \end{aligned}$$

For any fixed k and $0 \leq \forall j \leq k$, set $t_{k,j} = \frac{2j-k}{k}t_0$. The inequality above yields,

$$\sum_{j=1}^{k}\|\gamma_n(t_{k,j}) - \gamma_n(t_{k,j-1})\| \leq \frac{\tau_M + 2\frac{r_0}{n}}{\tau_M - \frac{2t_0}{k}}2t_0,$$

so

$$Length(\gamma_n) = \limsup_k \sum_{j=1}^{k}\|\gamma_n(t_{k,j}) - \gamma_n(t_{k,j-1})\| \leq \left(1 + \frac{2r_0}{\tau_M n}\right)2t_0.$$

Moreover, the γ_n's are curves joining q_1 to q_2 with images $\gamma_n([-t_0, t_0]) \subset \mathbb{R}^D \setminus \mathring{\mathcal{B}}(z_0, \tau_M)$, so that their lengths are at most that of the arc of great circle $c_{z_0}(q_1, q_2)$:

$$Length\left(\gamma_n\right) \geq Length\left(c_{z_0}(q_1, q_2)\right) = 2t_0.$$

Hence,

$$\lim_{n\to\infty} Length(\gamma_n) = 2t_0 = \tau_M\theta.$$

□

Lemma C.9. *Let M be a compact manifold, and $q_1, q_2 \in M$ with $q_1 \neq q_2$. Let $\{\gamma_n\}_{n\in\mathbb{N}}$ be a sequence of curves on M joining q_1 and q_2 such that $\sup_n Length(\gamma_n) < \infty$ Then there exists a curve γ on M joining q_1 and q_2 such that*

$$\liminf_n Length(\gamma_n) \leq Length(\gamma) \leq \limsup_n Length(\gamma_n).$$

Proof of Lemma C.9. Without loss of generality, take the γ_n's to be arc length parametrized. For all $n \in \mathbb{N}$, we let $g_n : [0,1] \to M$ be the reparametrization $g_n(t) = \gamma_n\left(Length(\gamma_n)t\right)$. Notice that for all $t \in [0,1]$, the set $\{g_n(t)\}_{n\in\mathbb{N}}$ is contained in the compact set M, so that it is bounded uniformly in t. Moreover, writing $K = \sup_n Length(\gamma_n) < \infty$, we have that for all $t_1, t_2 \in [0,1]$,

$$\begin{aligned}\|g_n(t_1) - g_n(t_2)\| &= \|\gamma_n\left(Length(\gamma_n)t_1\right) - \gamma_n\left(Length(\gamma_n)t_2\right)\| \\ &\leq Length(\gamma_n)|t_1 - t_2| \\ &\leq K|t_1 - t_2|.\end{aligned}$$

Hence, the sequence $\{g_n\}_{n\in\mathbb{N}}$ is pointwise bounded and equicontinuous. From Arzelà-Ascoli theorem [Mun75, Theorem 45.4], there exists a curve $\gamma : [0,1] \to M$ and subsequence $\{g_{n_i}\}_{i\in\mathbb{N}}$ converging uniformly to γ.

For any fixed k and $1 \leq \forall j \leq k$, set $t_{k,j} = \frac{j}{k}t_0$. The uniform convergence ensures that

$$\sum_{j=0}^{k} \|\gamma(t_{k,j+1}) - \gamma(t_{k,j})\| = \lim_{i\to\infty} \sum_{j=0}^{k} \|g_{n_i}(t_{k,j+1}) - g_{n_i}(t_{k,j})\|.$$

As a consequence,

$$\begin{aligned}Length(\gamma) &= \lim_{k\to\infty} \sum_{j=0}^{k} \|\gamma(t_{k,j+1}) - \gamma(t_{k,j})\| \\ &= \lim_{k\to\infty} \lim_{i\to\infty} \sum_{j=0}^{k} \|\gamma_{n_i}(t_{k,j+1}) - \gamma_{n_i}(t_{k,j})\| \\ &= \lim_{i\to\infty} Length(\gamma_{n_i}).\end{aligned}$$

Hence the result. □

Proof of Lemma V.15. Combining Lemma C.3 and Lemma C.9 provides the existence of a curve $\gamma \subset M$ joining q_1 and q_2 such that $Length(\gamma) = Length(c_{z_0}(q_1, q_2))$. But $M \subset \mathbb{R}^D \setminus \mathring{\mathcal{B}}(z_0, \tau_M)$, and since $\|q_1 - q_2\| < 2\tau_M$, $c_{z_0}(q_1, q_2)$ is the unique minimizing geodesic of $\partial\mathcal{B}(z_0, \tau_M) \subset \mathbb{R}^D \setminus \mathring{\mathcal{B}}(z_0, \tau_M)$ joining q_1 and q_2. Therefore, $\gamma = c_{z_0}(q_1, q_2) \subset M$, hence the result. □

Lemma C.10. *Let $M \in \mathcal{C}^{(3)}_{\tau_{min},L}$ be a submanifold with reach τ_M. For all $p \in M$, let us denote*

$$L_p := \sup_{q\in\mathcal{B}_M(p,\frac{\tau_M}{2}), v\in\mathcal{B}_{T_pM}(0,1)} \left\|\gamma'''_{q,v}(0)\right\|.$$

Then for all $r \leq \tau_M/2$,

$$\left| \sup_{v\in T_pM, \|v\|=1} \left\|\gamma''_{p,v}(0)\right\| - \sup_{q\in\mathcal{B}(p,r)\cap M} \frac{2d(q-p, T_pM)}{\|q-p\|^2} \right| \leq 3\left(\frac{1}{\tau_M^2} + L_p\right) r.$$

To prove Lemma C.10 we need the following straightforward result.

Lemma C.11. *Let U be a vector space and $u \in U$, $n \in U^\perp$. If $v = u + n + e$, then*

$$|d(v, U) - \|v - u\|\,| \leq \|e\|\,.$$

Proof of Lemma C.10. First note that for all unit vector $v \in T_pM$, $q_{v,r} = \gamma_{p,v}(r)$ belongs to $\mathcal{B}(p, r) \cap M$, and $r \leq \frac{\tau_M}{2}$ with Proposition C.1 (ii) implies $q_{v,r} \neq p$. Therefore, it suffices to show that for all $q \in \mathcal{B}(p, r) \cap M$, there exists $v = v_q \in T_pM$ such that

$$\left| \left\| \gamma''_{p,v}(0) \right\| - \frac{2d(q-p, T_pM)}{\|q-p\|^2} \right| \leq 3\left(\frac{1}{\tau_M^2} + L_p \right) r.$$

Let $q \in \mathcal{B}(p, r) \cap M$ be different from p. Denoting $t = d_M(p, q) > 0$, we call $\gamma = \gamma_{p,v}$ the arc-length parametrized geodesic of minimal length such that $\gamma(0) = p$ and $\gamma(t) = q$. γ exists from Proposition C.1 (ii), since $r \leq \frac{\tau_M}{2}$. A Taylor expansion at zero of γ yields,

$$\left\| \frac{q-p}{t} - \gamma'(0) - \frac{t}{2}\gamma''(0) \right\| \leq L_p \frac{t^2}{6}.$$

Since $\gamma''(0) \in T_pM^\perp$, Lemma C.11 shows that

$$\left| d(\frac{q-p}{t}, T_pM) - \left\| \frac{q-p}{t} - \gamma'(0) \right\| \right| \leq L_p \frac{t^2}{6}.$$

Therefore,

$$\begin{aligned}
&\left| \frac{2}{t} d(\frac{q-p}{t}, T_pM) - \|\gamma''(0)\| \right| \\
&\leq \frac{2}{t}\left(\left| d(\frac{q-p}{t}, T_pM) - \left\| \frac{q-p}{t} - \gamma'(0) \right\| \right| + \left\| \frac{q-p}{t} - \gamma'(0) - \frac{t}{2}\gamma''(0) \right\| \right) \\
&\leq \frac{2}{3} L_p t.
\end{aligned}$$

This yields,

$$\left| \frac{2d(q-p, T_pM)}{\|q-p\|^2} - \|\gamma''(0)\| \right| \leq 2d(q-p, T_pM) \left| \frac{1}{d_M(p,q)^2} - \frac{1}{\|q-p\|^2} \right| + \frac{2}{3} L_p t.$$

Moreover, from $\|q-p\| \leq d_M(p,q)$ and Lemma III.21, we derive

$$\begin{aligned}
\|q-p\|^2 \leq d_M(p,q)^2 &\leq \tau_M^2 \left(1 - \sqrt{1 - \frac{2\|q-p\|}{\tau_M}} \right)^2 \\
&\leq \tau_M^2 \frac{\left(\frac{\|q-p\|}{\tau_M} \right)^2}{\left(1 - \frac{2\|q-p\|}{\tau_M} \right)^{3/2}} \\
&\leq \frac{\|q-p\|^2}{1 - 3\frac{\|q-p\|}{\tau_M}},
\end{aligned}$$

where the last two inequalities follow from elementary real analysis arguments. Therefore, we get $t \leq 2\|q-p\|$ and

$$\left| \frac{1}{d_M(p,q)^2} - \frac{1}{\|q-p\|^2} \right| \leq \frac{3}{\tau_M \|q-p\|}.$$

Using moreover that $2d(q-p,T_pM) \leq \|q-p\|^2/\tau_M$ we derive,

$$\begin{aligned}\left|\|\gamma''(0)\| - \frac{2d(q-p,T_pM)}{\|q-p\|^2}\right| &\leq 2d(q-p,T_pM)\frac{3}{\tau_M\|q-p\|} + \frac{4}{3}L_p\|q-p\| \\ &\leq \frac{3}{\tau_M^2}\|q-p\| + \frac{4}{3}L_p\|q-p\| \\ &\leq 3\left(\frac{1}{\tau_M^2}+L_p\right)r.\end{aligned}$$

□

Proof of Lemma V.16. For $r>0$, let $\Delta_r := \{(p,q)\in M^2 | \, \|p-q\| < r\}$, and $\bar{\Delta} = \cap_{r>0}\Delta_r$ denote the diagonal of M^2. Consider the map $\varphi : M^2 \setminus \bar{\Delta} \to \mathbb{R}$ defined by $\varphi(p,q) = \frac{\|q-p\|^2}{2d(q-p,T_pM)}$. From (V.5), if there exists $p\neq q\in M$ such that $\tau_M = \varphi(p,q)$, then there exists $z\in Med(M)$ with $d(z,M)=\tau_M$. Hence, for all $p\neq q\in T_pM$, $\varphi(p,q)<\tau_M$, and by compactness of $M^2\backslash\Delta_r$, we have $\sup_{M^2\backslash\Delta_r}\varphi < \tau_M^{-1}$. Since we have the decomposition

$$\begin{aligned}\frac{1}{\tau_M} &= \sup_{(p,q)\in M^2\backslash\bar{\Delta}} \varphi(p,q) \\ &= \sup_{(p,q)\in M^2\backslash\Delta_r}\varphi(p,q) \vee \sup_{(p,q)\in\Delta_r\backslash\bar{\Delta}}\varphi(p,q),\end{aligned}$$

we get $\sup_{\Delta_r\backslash\bar{\Delta}}\varphi = \tau_M^{-1}$. Moreover, Lemma C.10 implies that

$$\left|\sup_{p\in M, v\in\mathcal{B}_{T_pM}(0,1)}\left\|\gamma''_{p,v}(0)\right\| - \sup_{(p,q)\in\Delta_r\backslash\bar{\Delta}}\varphi(p,q)\right| \leq 3\left(\frac{1}{\tau_M^2}+L\right)r$$

for $r>0$ small enough. Letting r go to zero yields

$$\sup_{p\in M, v\in\mathcal{B}_{T_pM}(0,1)}\left\|\gamma''_{p,v}(0)\right\| = \frac{1}{\tau_M}.$$

Finally, the unit tangent bundle $T^{\leq 1}M = \{(p,v), p\in M, v\in\mathcal{B}_{T_pM}(0,1)\}$ being compact, there exists (q_0,v_0) such that $\gamma_0 = \gamma_{p_0,v_0}$ satisfies $\|\gamma_0''(0)\| = \tau_M^{-1}$, which concludes the proof. □

C.3 Analysis of the Estimator

C.3.1 Global Case

To show Proposition V.20, we show a stronger result (Proposition C.12) that applies to a reach attaining pair with any size 2λ, meaning that it is not necessarily a bottleneck. Proposition V.20 follows straightforwardly by setting λ equal to τ_M.

Proposition C.12. *Let $M\subset\mathbb{R}^D$ be a submanifold, and $0<\lambda\leq\tau_M$. Assume that M has a reach attaining pair $(q_1,q_2)\in M^2$ (see Definition V.14) with $\|q_1-q_2\|\geq 2\lambda$. Let $\mathcal{X}\subset M$. If there exists $x,y\in\mathcal{X}$ with $\|q_1-x\|<\lambda$ and $\|q_2-y\|<\lambda$, then*

$$0 \leq \frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathcal{X})} \leq \frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\{x,y\})} \leq C_{\tau_M,\lambda}\max\left\{d_M(q_1,x), d_M(q_2,y)\right\},$$

where $C_{\tau_M,\lambda} = \frac{2\tau_M^2+6\tau_M\lambda+\lambda^2}{2\tau_M^2\lambda^2}$ depends only on the parameters τ_M,λ, and is a decreasing function of τ_M and λ when the other parameter is fixed.

Proof of Proposition C.12. The two left hand inequalities are a direct consequence of Corollary V.19, let us then focus on the third one.

Without loss of generality, assume that $\|q_1 - q_2\| = 2\lambda$. We set $t := \max\{d_M(q_1, x), d_M(q_2, y)\}$ and $z_1 := x + (q_2 - q_1)$. We have $\|z_1 - x\| = \|q_2 - q_1\| = 2\lambda$ and $\|y - q_2\|, \|q_1 - x\| \leq t$. Therefore, from the definition of $\hat{\tau}$ in (V.18) and the fact that the distance function to a linear space is 1-Lipschitz, we get

$$\begin{aligned}\frac{1}{\hat{\tau}(\{x,y\})} &\geq \frac{2d(y-x, T_xM)}{\|y-x\|^2} \\ &= \frac{2d\left((y-q_2)+(z_1-x)+(q_1-x), T_xM\right)}{\|(y-q_2)+(z_1-x)+(q_1-x)\|^2} \\ &\geq \frac{d(z_1-x, T_xM) - 2t}{2(\lambda+t)^2}.\end{aligned}$$

Let now $\theta := \angle(q_2 - q_1, T_{q_1}M) = \min_{v \in T_{q_1}M} \angle(q_2 - q_1, v)$. Since $z_0 \in Med(M)$, with $q_1, q_2 \in \mathcal{B}(z_0, \tau_M)$ and $\|q_1 - q_2\| = 2\lambda$, for any v' such that $v' \perp z_0 - q_1$, we have $\angle(q_2 - q_1, v') \geq \frac{\pi}{2} - \angle(q_2 - q_1, z_0 - q_1)$. Hence, $\sin\theta \geq \frac{\lambda}{\tau_M}$ and $\cos\theta \leq \frac{\sqrt{\tau_M^2 - \lambda^2}}{\tau_M}$. Let $v_1 \in T_{q_1}M$ be any point in $T_{q_1}M$ realizing this angle, in the sense that $\angle(q_2 - q_1, v_1) = \angle(q_2 - q_1, T_{q_1}M)$. Then we have

$$\angle(z_1 - x, v_1) = \angle(q_2 - q_1, v_1) = \theta.$$

Let $\bar{v}_1 \in T_xM$ be the parallel transport of v_1 along the geodesic between q_1 and x. Since M has reach τ_M, Proposition C.1 (vi) gives

$$\angle(v_1, \bar{v}_1) \leq \frac{d_M(x, q_1)}{\tau_M} \leq \frac{t}{\tau_M}.$$

Hence the angle $\angle(z_1 - x, T_xM)$ can be lower bounded as

$$\begin{aligned}\angle(z_1 - x, T_xM) &\geq \angle(z_1 - x, \bar{v}_1) \\ &\geq \angle(z_1 - x, v) - \angle(v, \bar{v}_1) \\ &\geq \theta - \frac{t}{\tau_M}.\end{aligned}$$

And $0 \leq \frac{\lambda}{\tau_M} - \frac{t}{\tau_M} \leq \theta - \frac{t}{\tau_M} \leq \angle(z_1 - x, T_xM) \leq \frac{\pi}{2}$, so the inequality is preserved by the sine function, i.e.

$$\begin{aligned}d(z_1 - x, T_xM) &= \|z_1 - x\| \sin(\angle(z_1 - x, T_xM)) \\ &\geq 2\lambda \sin\left(\theta - \frac{t}{\tau_M}\right) = 2\lambda\left(\sin\theta\cos\frac{t}{\tau_M} - \cos\theta\sin\frac{t}{\tau_M}\right) \\ &= \frac{2\lambda^2}{\tau_M}\cos\frac{t}{\tau_M} - \frac{2\lambda\sqrt{\tau_M^2 - \lambda^2}}{\tau_M}\sin\frac{t}{\tau_M}.\end{aligned}$$

Combining the previous bounds yields,

$$\begin{aligned}\frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\{x,y\})} &\leq \frac{1}{\tau_M} - \frac{d(z_1 - x, T_xM) - 2t}{2(\lambda+t)^2} \\ &\leq \frac{1}{\tau_M} - \frac{\frac{1}{\tau_M}\cos\frac{t}{\tau_M} - \frac{\sqrt{\tau_M^2-\lambda^2}}{\tau_M\lambda}\sin\frac{t}{\tau_M} - \frac{t}{\lambda^2}}{\left(1 + \frac{t}{\lambda}\right)^2}.\end{aligned}$$

Using again that $t < \lambda \leq \tau_M$, the latter right-hand side term is itself upper bounded by,

$$\begin{aligned}
&\frac{1}{\tau_M} - \left(\frac{1}{\tau_M}\left(1 - \frac{t^2}{2\tau_M^2}\right) - \frac{\sqrt{\tau_M^2 - \lambda^2}}{\tau_M \lambda}\frac{t}{\tau_M} - \frac{t}{\lambda^2}\right)\left(1 - \frac{2t}{\lambda}\right) \\
&\quad \leq \left(\frac{\lambda}{2\tau_M^3} + \frac{\sqrt{\tau_M^2 - \lambda^2}}{\tau_M^2 \lambda} + \frac{1}{\lambda^2} + \frac{2}{\lambda \tau_M}\right) t \\
&\quad = \frac{2\tau_M^3 + 2\lambda\tau_M\sqrt{\tau_M^2 - \lambda^2} + 4\tau_M^2\lambda + \lambda^3}{2\tau_M^3\lambda^2} t \\
&\quad \leq \frac{2\tau_M^2 + 6\tau_M\lambda + \lambda^2}{2\tau_M^2\lambda^2} t := C_{\tau_M,\lambda} t,
\end{aligned}$$

which is the announced result. □

As for Proposition V.20, we tackle the proof of Proposition V.21 by showing the following stronger one, Proposition C.13. Proposition V.21 follows straightforwardly by setting λ equal to τ_M.

Proposition C.13. *Let $P \in \mathcal{P}^{(3)}_{\tau_{\min},L}(f_{\min})$, $M = Supp(P)$ and $0 < \lambda \leq \tau_M$. Assume that M has a reach attaining pair $(q_1, q_2) \in M^2$ (see Definition V.14) with $\|q_1 - q_2\| \geq 2\lambda$. Then*

$$\mathbb{E}_{P^n}\left[\left|\frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathbb{X}_n)}\right|^p\right] \leq C_{\tau_M,\lambda,f_{min},d,p} n^{-\frac{p}{d}}.$$

where $C_{\tau_M,\lambda,f_{min},d,p}$ depends only on τ_M, λ, f_{min} d, p, and is a decreasing function of τ_M and λ when other parameters are fixed.

Proof of Proposition C.13. Let $s < \frac{1}{\tau_M}$, $C_{\tau_M,\lambda} = \frac{2\tau_M^2 + 6\tau_M\lambda + \lambda^2}{2\tau_M^2\lambda^2}$, and $t = \frac{1}{C_{\tau_M,\lambda}} s \leq 2\tau_M/9$. Let $\omega_d := \mathcal{H}^d(\mathcal{B}_{\mathbb{R}^d}(0,1))$ be the volume of the d-dimensional unit ball. Then note that from Proposition C.1 (v), for all $q \in M$,

$$\begin{aligned}
P\left(\mathcal{B}_M(p,t)\right) &\geq f_{min}\mathcal{H}^d\left(\mathcal{B}_M(p,t)\right) \\
&\geq \omega_d f_{min}\left(1 - \left(\frac{t}{6\tau_M}\right)^2\right)^d t^d \\
&\geq \omega_d f_{min}\left(\frac{728}{729}\right)^d t^d.
\end{aligned}$$

Moreover, Proposition V.20 asserts that $\left|\frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathbb{X}_n)}\right| > s$ implies that either $\mathcal{B}_M(q_1,t) \cap \mathbb{X}_n = \emptyset$ or $\mathcal{B}_M(q_2,t) \cap \mathbb{X}_n = \emptyset$. Hence,

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathbb{X}_n)}\right| > s\right) &\leq \mathbb{P}\left(\mathcal{B}_M(q_1,t) \cap \mathbb{X}_n = \emptyset\right) + \mathbb{P}\left(\mathcal{B}_M(q_2,t) \cap \mathbb{X}_n = \emptyset\right) \\
&\leq 2\left(1 - \omega_d f_{min}\left(\frac{728}{729}\right)^d t^d\right)^n \\
&\leq 2\exp\left(-n\omega_d f_{min}\left(\frac{728}{729}\right)^d C^{-d}_{\tau_M,\lambda} s^d\right).
\end{aligned}$$

Letting $\Gamma(\cdot)$ denote the Gamma function, the integration of the above bound gives

$$\begin{aligned}
\mathbb{E}_{P^n}\left[\left|\frac{1}{\tau_M}-\frac{1}{\hat{\tau}(\mathbb{X}_n)}\right|^p\right] &= \int_0^{\frac{1}{\tau_M^p}} \mathbb{P}\left(\left|\frac{1}{\tau_M}-\frac{1}{\hat{\tau}(\mathbb{X}_n)}\right|^p > s\right) ds \\
&\leq 2\int_0^{\infty} \exp\left(-n\omega_d f_{min}\left(\frac{728}{729}\right)^d C_{\tau_M,\lambda}^{-d} s^{\frac{d}{p}}\right) ds \\
&= \frac{2\left(\frac{729}{728}\right)^{\frac{p}{d}} C_{\tau_M,\lambda}^{p}}{\left(n\omega_d f_{min}\right)^{\frac{p}{d}}} \int_0^{\infty} x^{\frac{p}{d}-1} e^{-x} dx \\
&= \frac{2\left(\frac{729}{728}\right)^{\frac{p}{d}} \Gamma\left(\frac{p}{d}\right)}{\left(\omega_d f_{min}\right)^{\frac{p}{d}}} C_{\tau_M,\lambda} n^{-\frac{p}{d}} \\
&:= C_{\tau_M,\lambda,f_{\min},d,p} n^{-\frac{p}{d}}.
\end{aligned}$$

where $C_{\tau_M,\lambda,f_{\min},d,p}$ depends only on τ_M, λ, f_{min}, d, p, and is a decreasing function of τ_M and λ when other parameters are fixed. □

C.3.2 Local Case

Lemma C.14. *Let M be a submanifold and $p \in M$. Let $v_0, v_1 \in T_pM$ be a unit tangent vector, and let $\theta = \angle(v_0, v_1)$. Let $\gamma_{p,v}$ be the arc length parametrized geodesic starting from p with velocity v, and write $\gamma_i = \gamma_{p,v_i}$ for $i = 0, 1$. Let $\kappa_p = \max_{v\in\mathcal{B}_{T_pM}(0,1)} \left\|\gamma''_{q_0,v}(0)\right\|$. Then,*

$$\|\gamma_1''(0)\| \geq \|\gamma_0''(0)\| - \frac{\sqrt{2}}{\sqrt{2}-1}\sin^2\theta\left(\kappa_p + \|\gamma_0''(0)\|\right) - \frac{1}{\sqrt{2}-1}\left(\kappa_p - \|\gamma_0''(0)\|\right). \tag{C.15}$$

and

$$\begin{aligned}
\|\gamma_1''(0)\| \geq{}& \|\gamma_0''(0)\| - \sin^2\theta\left(\kappa_p + \|\gamma_0''(0)\|\right) \\
&- \frac{|\cos\theta\sin\theta|\,\kappa_p\sqrt{\kappa_p - \|\gamma_0''(0)\|}}{(\sqrt{2}-1)\,\|\gamma_0''(0)\|}\left(\frac{2\kappa_p}{\|\gamma_0''(0)\|}+1\right).
\end{aligned} \tag{C.16}$$

Proof of Lemma C.14. Let $w \in T_pM$ be a unit vector satisfying $w \perp v_0$ and $v_1 = \cos\theta v_0 + \sin\theta w$. For $t \in \mathbb{R}$, let $v(t) := (\cos t)v_0 + (\sin t)w \in T_pM$, so that $v_1 = v(\theta)$. Then

$$\begin{aligned}
\left\|d_0^2\exp_p(v(t),v(t))\right\| &= \Big\|\cos^2 t d_0^2\exp_p(v_0,v_0) + 2\cos t\sin t d_0^2\exp_p(v_0,w) \\
&\quad + \sin^2 t d_0^2\exp_p(w,w)\Big\| \\
&\geq |\cos t|\left\|\cos t d_0^2\exp_p(v_0,v_0) + 2\sin t d_0^2\exp_p(v_0,w)\right\| \\
&\quad - \sin^2 t\left\|d_0^2\exp_p(w,w)\right\|.
\end{aligned} \tag{C.17}$$

Now, note that when $x \in [-1,1]$, $\sqrt{1+x} \geq 1 + f(x)$, where $f(x) = \min\{x, (\sqrt{2}-1)x\}$. Hence for any $v', v'' \in T_pM$,

$$\begin{aligned}
\|v'+v''\| &= \sqrt{\|v'\|^2+\|v''\|^2}\sqrt{1+\frac{2\langle v', v''\rangle}{\|v'\|^2+\|v''\|^2}} \\
&\geq \sqrt{\|v'\|^2+\|v''\|^2}\left(1+f\left(\frac{2\langle v', v''\rangle}{\|v'\|^2+\|v''\|^2}\right)\right) \\
&\geq \|v'\| + f\left(\frac{2\langle v', v''\rangle}{\sqrt{\|v'\|^2+\|v''\|^2}}\right).
\end{aligned}$$

Applying the latter inequality to (C.17) and using $d_0^2 \exp_p(v_0, v_0) = \gamma_0''(0)$ and $d_0^2 \exp_p(w, w) \leq \kappa_p$ gives

$$\begin{aligned}
&\left\| d_0^2 \exp_p(v(t), v(t)) \right\| \\
&\geq \cos^2 t \left\| d_0^2 \exp_p(v_0, v_0) \right\| - \sin^2 t \left\| d_0^2 \exp_p(w, w) \right\| \\
&\quad + |\cos t| f \left(\frac{4 \cos t \sin t \left\langle d_0 \exp_p(v_0, v_0), d_0 \exp_p(v_0, w) \right\rangle}{\sqrt{\cos^2 t \left\| d_0^2 \exp_p(v_0, v_0) \right\|^2 + 4 \sin^2 t \left\| d_0^2 \exp_p(v_0, w) \right\|^2}} \right) \\
&\geq \cos^2 t \left\| \gamma_0''(0) \right\| - \kappa_p \sin^2 t \\
&\quad + |\cos t| f \left(\frac{4 \cos t \sin t \left\langle \gamma_0''(0), d_0 \exp_p(v_0, w) \right\rangle}{\sqrt{\cos^2 t \left\| \gamma_0''(0) \right\|^2 + 4 \sin^2 t \left\| d_0^2 \exp_p(v_0, w) \right\|^2}} \right).
\end{aligned}$$

Now, note that $f(x) \geq -|x|$ for $x \in [-1, 1]$, so applying this with $t = \theta$ gives

$$\begin{aligned}
\left\| \gamma_1''(0) \right\| &= \left\| d_0^2 \exp_p(v_1, v_1) \right\| \\
&\geq \cos^2 \theta \left\| \gamma_0''(0) \right\| - \sin^2 \theta \kappa_p \\
&\quad - \frac{4 \left| \cos^2 \theta \sin \theta \left\langle \gamma_0''(0), d_0 \exp_p(v_0, w) \right\rangle \right|}{\sqrt{\cos^2 \theta \left\| \gamma_0''(0) \right\|^2 + 4 \sin^2 \theta \left\| d_0^2 \exp_p(v_0, w) \right\|^2}}.
\end{aligned} \tag{C.18}$$

We now focus on the third term of the right-hand side. For this, note that either

$$t \sin t \langle \gamma_0''(0), d_0 \exp_p(v_0, w) \rangle \geq 0,$$

or

$$\cos(-t) \sin(-t) \langle \gamma_0''(0), d_0 \exp_p(v_0, w) \rangle \geq 0,$$

so that

$$\begin{aligned}
\kappa_p &\geq \max \left\{ \left\| d_0^2 \exp_p(v(-t), v(-t)) \right\|, \left\| d_0^2 \exp_p(v(t), v(t)) \right\| \right\} \\
&\geq \cos^2 t \left\| \gamma_0''(0) \right\| + \frac{4(\sqrt{2} - 1) \left| \cos^2 t \sin t \left\langle \gamma_0''(0), d_0 \exp_p(v_0, w) \right\rangle \right|}{\sqrt{\cos^2 t \left\| \gamma_0''(0) \right\|^2 + 4 \sin^2 t \left\| d_0^2 \exp_p(v_0, w) \right\|^2}} \\
&\quad - \sin^2 t \kappa_p.
\end{aligned}$$

As a consequence,

$$\begin{aligned}
&\frac{\left| \cos^2 t \sin t \left\langle \gamma_0''(0), d_0 \exp_p(v_0, w) \right\rangle \right|}{\sqrt{\cos^2 t \left\| \gamma_0''(0) \right\|^2 + 4 \sin^2 t \left\| d_0^2 \exp_p(v_0, w) \right\|^2}} \\
&\leq \frac{1}{4(\sqrt{2} - 1)} \left((1 + \sin^2 t) \kappa_p - \cos^2 t \left\| \gamma_0''(0) \right\| \right) \\
&= \frac{1}{4(\sqrt{2} - 1)} \left(\cos^2 t \left(\kappa_p - \left\| \gamma_0''(0) \right\| \right) + 2 \sin^2 t \kappa_p \right).
\end{aligned}$$

First, setting $t = \theta$, we derive

$$\begin{aligned}
&\|\gamma_1''(0)\| \\
&\geq \cos^2\theta \,\|\gamma_0''(0)\| - \left(1 + \frac{2}{\sqrt{2}-1}\right)\sin^2\theta\kappa_p - \frac{1}{\sqrt{2}-1}\cos^2\theta\left(\kappa_p - \|\gamma_0''(0)\|\right) \\
&= \|\gamma_0''(0)\| - \frac{\sqrt{2}}{\sqrt{2}-1}\sin^2\theta\left(\kappa_p + \|\gamma_0''(0)\|\right) - \frac{1}{\sqrt{2}-1}\left(\kappa_p - \|\gamma_0''(0)\|\right).
\end{aligned}$$

Furthermore, let t_0 be defined by $\sin^2 t_0 = 1 - \frac{\|\gamma_0''(0)\|}{\kappa_p} + \epsilon$ for $\epsilon > 0$ small enough. Then $\sqrt{\cos^2 t_0 \|\gamma_0''(0)\|^2 + 4\sin^2 t_0 \left\|d_0^2 \exp_p(v_0, w)\right\|^2} \leq \kappa_p$, yielding

$$\begin{aligned}
&\left|\left\langle \gamma_0''(0), d_0 \exp_p(v_0, w)\right\rangle\right| \\
&\leq \frac{\sqrt{\kappa_p}}{4(\sqrt{2}-1)\cos^2 t_0 |\sin t_0|}\left(\cos^2 t_0\left(\kappa_p - \|\gamma_0''(0)\|\right) + 2\sin^2 t_0\kappa_p\right) \\
&= \frac{\kappa_p^{\frac{3}{2}}}{4(\sqrt{2}-1)}\left(\frac{1 - \frac{\|\gamma_0''(0)\|}{\kappa_p}}{\sqrt{1 - \frac{\|\gamma_0''(0)\|}{\kappa_p} + \epsilon}} + \frac{2\sqrt{1 - \frac{\|\gamma_0''(0)\|}{\kappa_p} + \epsilon}}{\frac{\|\gamma_0''(0)\|}{\kappa_p} - \epsilon}\right).
\end{aligned}$$

Sending $\epsilon \to 0$, we obtain

$$\left|\left\langle \gamma_0''(0), d_0 \exp_p(v_0, w)\right\rangle\right| \leq \frac{\kappa_p\sqrt{\kappa_p - \|\gamma_0''(0)\|}}{4(\sqrt{2}-1)}\left(\frac{2\kappa_p}{\|\gamma_0''(0)\|} + 1\right).$$

Using the previous bound together with

$$\cos^2\theta\,\|\gamma_0''(0)\|^2 + 4\sin^2\theta\left\|d_0^2\exp_p(v_0, w)\right\|^2 \geq |\cos\theta|\,\|\gamma_0''(0)\|,$$

we finally obtain

$$\begin{aligned}
\|\gamma_1''(0)\| \geq \|\gamma_0''(0)\| &- \sin^2\theta\left(\kappa_p + \|\gamma_0''(0)\|\right) \\
&- \frac{|\cos\theta\sin\theta|\,\kappa_p\sqrt{\kappa_p - \|\gamma_0''(0)\|}}{(\sqrt{2}-1)\,\|\gamma_0''(0)\|}\left(\frac{2\kappa_p}{\|\gamma_0''(0)\|} + 1\right).
\end{aligned}$$

□

Proof of Lemma V.22. Note first from Proposition C.1 (ii), $d_M(x, y) < \pi\tau_M$ ensures the existence and uniqueness of the geodesic $\gamma_{x\to y}$. The two left hand inequalities are a direct consequence of Corollary V.19. Let us then focus on the third one. Let $t_0 := d_M(x, y)$, and write $\gamma = \gamma_{x\to y}$ for short. By definition of $\hat{\tau}$ in (V.18),

$$\frac{1}{\hat{\tau}(\{x, y\})} \geq \frac{2d(y - x, T_xM)}{\|y - x\|^2} \geq \frac{2d(y - x, T_xM)}{t_0^2}. \tag{C.19}$$

Let $H_{\gamma''(0)} := \{x + u \in \mathbb{R}^D | \left\langle u, \gamma''_{x\to y}(0)\right\rangle = 0\}$ denote the affine hyperplane passing though x with the normal vector $\gamma''(0)$. Since $\gamma''(0) \in T_xM^\perp$, $T_xM \subset H_{\gamma''(0)}$. As a consequence,

$$d\left(y - x, T_xM\right) \geq d\left(y - x, H_{\gamma''(0)}\right) = \frac{|\langle\gamma''(0), y - x\rangle|}{\|\gamma''(0)\|}. \tag{C.20}$$

Using the Taylor expansion of γ at order two, we get

$$y - x = \gamma(t_0) - \gamma(0) = t_0\gamma'(0) + \int_0^{t_0}\int_0^t \gamma''(s)dsdt. \tag{C.21}$$

Since γ is parametrized by arc length, $\langle\gamma'(t), \gamma'(t)\rangle = 1$. Differentiating this identity at 0 yields $\langle\gamma''(0), \gamma'(0)\rangle = 0$. In addition, by definition of $\mathcal{C}^{(3)}_{\tau_{\min},L} \ni M$ (Definition V.7), the geodesic γ satisfies $\|\gamma''(s) - \gamma''(0)\| \leq L|s|$. Therefore,

$$\begin{aligned}|\langle\gamma''(0), \gamma''(s)\rangle| &= |\langle\gamma''(0), \gamma''(0)\rangle - \langle\gamma''(0), \gamma''(s) - \gamma''(0)\rangle| \\ &\geq ||\gamma''(0)||^2 - L||\gamma''(0)||s|.\end{aligned}$$

Combining the above bound together with (C.19), (C.20) and (C.21), we derive

$$\frac{1}{\hat{\tau}(\{x, y\})} \geq \|\gamma''(0)\| - \frac{2}{3}Lt_0,$$

which is the announced inequality. □

Proof of Lemma V.23. For short, in what follows, we let $t_x := d_M(q_0, x)$, $t_y := d_M(q_0, y)$, and $\theta := \angle(\gamma'_{x\to y}(0), \gamma'_{q_0\to x}(t_x))$. From Lemma C.14,

$$\begin{aligned}\left\|\gamma''_{x\to y}(0)\right\| &\geq \left\|\gamma''_{q_0\to x}(t_x)\right\| - \frac{\sqrt{2}}{\sqrt{2}-1}\sin^2\theta\left(\kappa_x + \left\|\gamma''_{q_0\to x}(t_x)\right\|\right) \\ &\quad - \frac{1}{\sqrt{2}-1}\left(\kappa_x - \left\|\gamma''_{q_0\to x}(t_x)\right\|\right) \\ &= \frac{\sqrt{2}}{\sqrt{2}-1}\cos^2\theta\left\|\gamma''_{q_0\to x}(t_x)\right\| - \left(\frac{1}{\sqrt{2}-1} + \frac{\sqrt{2}}{\sqrt{2}-1}\sin^2\theta\right)\kappa_x. \end{aligned}\tag{C.22}$$

We now focus on the term $\left\|\gamma''_{q_0\to x}(t_x)\right\|$. Applying again Lemma C.14 yields

$$\left\|\gamma''_{q_0\to x}(0)\right\| \geq (1 - 2\sin^2\theta_x)\kappa_{q_0},$$

and since $\gamma''_{q_0\to x}$ is L-Lipschitz,

$$\begin{aligned}\left\|\gamma''_{q_0\to x}(t_x)\right\| &\geq \left\|\gamma''_{q_0\to x}(0)\right\| - \left\|\gamma''_{q_0\to x}(t_x) - \gamma''_{q_0\to x}(0)\right\| \\ &\geq (1 - 2\sin^2\theta_x)\kappa_{q_0} - Lt_x. \end{aligned}\tag{C.23}$$

Now we focus on bounding the terms $\sin^2\theta$ and $\cos^2\theta$. Let $\mathcal{S}^2_{\tau_M}$ be a d-dimensional sphere of radius τ_M. In what follows, for short, $\angle abc$ stands for $\angle(\gamma'_{b\to a}(0), \gamma'_{b\to c}(0))$. First, let $\tilde{q}_0, \tilde{x}, \tilde{y} \in \mathcal{S}^2_{\tau_M}$ be such that $d_{\mathcal{S}^2_{\tau_M}}(\tilde{q}_0, \tilde{x}) = d_M(q_0, x)$, $d_{\mathcal{S}^2_{\tau_M}}(\tilde{q}_0, \tilde{y}) = d_M(q_0, y)$, and $\angle\tilde{x}\tilde{q}_0\tilde{y} = \angle xq_0y$. Then from Toponogov's comparison Theorem [Mey89], we have $d_{\mathcal{S}^2_{\tau_M}}(\tilde{x}, \tilde{y}) \leq d_M(x, y)$. Moreover, the spherical law of cosines [Tod79] writes as

$$\cos\left(\frac{d_{\mathcal{S}^2_{\tau_M}}(\tilde{x}, \tilde{y})}{\tau_M}\right) = \cos\left(\frac{t_x}{\tau_M}\right)\cos\left(\frac{t_y}{\tau_M}\right) + \sin\left(\frac{t_x}{\tau_M}\right)\sin\left(\frac{t_y}{\tau_M}\right)\cos\left(\angle\tilde{x}\tilde{q}_0\tilde{y}\right),$$

and since $t_x, t_y \leq \frac{\pi}{2}$ and $\cos(\cdot)$ is decreasing on $[0, \pi]$, we get

$$t_y \leq d_{\mathcal{S}^2_{\tau_M}}(\tilde{x}, \tilde{y}) \leq d_M(x, y).$$

Now, let $\bar{q}_0, \bar{x}, \bar{y} \in \mathcal{S}^2_{\tau_M}$ be such that $d_{\mathcal{S}^2_{\tau_M}}(\bar{q}_0, \bar{x}) = d_M(q_0, x)$, $d_{\mathcal{S}^2_{\tau_M}}(\bar{q}_0, \bar{y}) = d_M(q_0, y)$, and $d_{\mathcal{S}^2_{\tau_M}}(\bar{x}, \bar{y}) = d_M(x, y)$. Applying Toponogov's comparison Theorem [Mey89], we have $\angle q_0 x y \leq \angle \bar{q}_0 \bar{x} \bar{y}$ and $\angle x q_0 y \leq \angle \bar{x} \bar{q}_0 \bar{y}$, and from the spherical law of cosines [Tod79],

$$\cos\left(\angle \bar{q}_0 \bar{x} \bar{y}\right) = \frac{\cos\left(\frac{t_y}{\tau_M}\right) - \cos\left(\frac{t_x}{\tau_M}\right)\cos\left(\frac{d_M(x,y)}{\tau_M}\right)}{\sin\left(\frac{t_x}{\tau_M}\right)\sin\left(\frac{d_M(x,y)}{\tau_M}\right)} \geq 0,$$

so that $\angle q_0 x y \leq \angle \bar{q}_0 \bar{x} \bar{y} \leq \frac{\pi}{2}$. Also, $\angle x q_0 y \geq |\theta_x - \theta_y| \geq \frac{\pi}{2}$ yields $\frac{\pi}{2} \leq \angle x q_0 y \leq \angle \bar{x} \bar{q}_0 \bar{y}$, and $\theta = \angle(\gamma'_{x \to y}(0), \gamma'_{q_0 \to x}(t_x)) = \pi - \angle q_0 x y$. Hence applying the spherical law of sines and cosines [Tod79] implies

$$\begin{aligned}
\sin\theta &= \sin(\angle q_0 x y) \leq \sin(\angle \bar{q}_0 \bar{x} \bar{y}) \\
&= \frac{\sin\left(\frac{t_y}{\tau_M}\right)\sin(\angle \bar{x} \bar{q}_0 \bar{y})}{\sqrt{1 - \left(\cos\left(\frac{t_x}{\tau_M}\right)\cos\left(\frac{t_y}{\tau_M}\right) + \sin\left(\frac{t_x}{\tau_M}\right)\sin\left(\frac{t_y}{\tau_M}\right)\cos(\angle \bar{x} \bar{q}_0 \bar{y})\right)^2}} \\
&\leq \frac{\sin\left(\frac{t_y}{\tau_M}\right)\sin(\angle \bar{x} \bar{q}_0 \bar{y})}{\sqrt{1 - \cos^2\left(\frac{t_x}{\tau_M}\right)\cos^2\left(\frac{t_y}{\tau_M}\right)}} \\
&= \frac{\sin\left(\frac{t_y}{\tau_M}\right)\sin(\angle \bar{x} \bar{q}_0 \bar{y})}{\sqrt{\sin^2\left(\frac{t_y}{\tau_M}\right) + \sin^2\left(\frac{t_x}{\tau_M}\right)\cos^2\left(\frac{t_y}{\tau_M}\right)}} \\
&\leq \sin(\angle \bar{x} \bar{q}_0 \bar{y}) \leq \sin(\angle x q_0 y) \leq \sin(|\theta_x - \theta_y|).
\end{aligned} \tag{C.24}$$

And accordingly,

$$|\cos\theta| = \sqrt{1 - \sin^2\theta} \geq \sqrt{1 - \sin^2(|\theta_x - \theta_y|)} = |\cos(|\theta_x - \theta_y|)|. \tag{C.25}$$

Hence applying (C.23), (C.24), and (C.25) to (C.22) gives

$$\begin{aligned}
&\left\|\gamma''_{x \to y}(0)\right\| \\
&\geq \frac{\sqrt{2}}{\sqrt{2}-1}\cos^2(|\theta_x - \theta_y|)\left((1 - 2\sin^2\theta_x)\kappa_{q_0} - L t_x\right) \\
&\quad - \left(\frac{1}{\sqrt{2}-1} + \frac{\sqrt{2}}{\sqrt{2}-1}\sin^2(|\theta_x - \theta_y|)\right)\kappa_x \\
&= \frac{(\sqrt{2}\kappa_{q_0} - \kappa_x)}{\sqrt{2}-1} \\
&\quad - \frac{\sqrt{2}}{\sqrt{2}-1}\left((\kappa_{q_0} + \kappa_x)\sin^2(|\theta_x - \theta_y|) + 2\kappa_{q_0}\sin^2\theta_x\cos^2(|\theta_x - \theta_y|)\right) \\
&\quad - \frac{\sqrt{2}}{\sqrt{2}-1} L t_x \cos^2(\theta_x + \theta_y) \\
&\geq \kappa_{q_0} - \frac{1}{\sqrt{2}-1}\left(\kappa_x - \kappa_{q_0} + \sqrt{2}(3\kappa_{q_0} + \kappa_x)\sin^2(|\theta_x - \theta_y|) + \sqrt{2} L t_x\right).
\end{aligned}$$

□

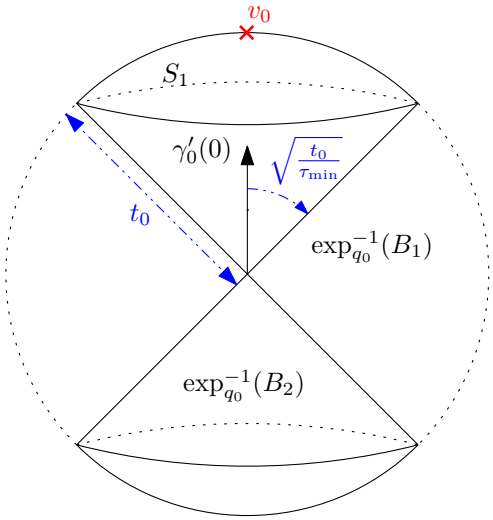


Figure C.2 – Layout of Proposition V.25.

Proof of Proposition V.25. In what follows, we let $t_0 \leq \frac{\tau_{\min}}{10}$,

$$B_1 := \exp_{q_0}\Big(\Big\{v \in T_{q_0}M : \|v\| \leq t_0,\, \angle(\gamma_0'(0), v) \leq \sqrt{\frac{t_0}{\tau_{\min}}}\Big\}\Big),$$

$$B_2 := \exp_{q_0}\Big(\Big\{v \in T_{q_0}M : \|v\| \leq t_0,\, \angle(\gamma_0'(0), v) \geq \pi - \sqrt{\frac{t_0}{\tau_{\min}}}\Big\}\Big),$$

and $B_0 := B_1 \cup B_2$, as in Figure C.2. Let $\mathcal{X} \subset M$, and $x, y \in \mathcal{X}$ be such that $x \in B_1$, $y \in B_2$. Writing $\theta_x := \angle(\gamma_0'(0), \gamma_{q_0 \to x}'(0))$ and $\theta_y := \angle(\gamma_0'(0), \gamma_{q_0 \to y}'(0))$, then $\theta_x \leq \sqrt{\frac{t_0}{\tau_{\min}}} \leq \frac{\pi}{4}$ and $\theta_y \geq \pi - \sqrt{\frac{t_0}{\tau_{\min}}} \geq \frac{3\pi}{4}$. Also, $d_M(q_0, x) \leq t_0$ and $d_M(x, y) \leq 2t_0$, so that

$$\begin{aligned}
0 &\leq \frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathcal{X})} \\
&\leq \frac{4\sqrt{2}\sin^2(|\theta_x - \theta_y|)}{(\sqrt{2}-1)\tau_M} + L\left(\frac{2}{3}d_M(x,y) + \frac{\sqrt{2}}{\sqrt{2}-1}d_M(q_0,x)\right) \\
&\leq \left(\frac{16\sqrt{2}}{(\sqrt{2}-1)\tau_{\min}\tau_M} + \frac{(7\sqrt{2}-4)L}{3(\sqrt{2}-1)}\right)t_0.
\end{aligned}$$

A symmetric argument also applies when $x \in B_2$ and $y \in B_1$. Now, for any $s < \frac{1}{\tau_M}$, let $t_0(s) := \left(\frac{16\sqrt{2}}{(\sqrt{2}-1)\tau_{\min}\tau_M} + \frac{(7\sqrt{2}-4)L}{3(\sqrt{2}-1)}\right)^{-1} s < \frac{\tau_{\min}}{10}$. The above argument implies that if $\frac{1}{\hat{\tau}(\mathcal{X})} < \frac{1}{\tau_M} - s$, then for any $x, y \in \mathcal{X} \cap B_0$, one has either $x, y \in B_1$ or $x, y \in B_2$. Hence

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{\tau_M} - \frac{1}{\hat{\tau}(X)} > s\right) \\
&\leq \sum_{m=0}^{n}\binom{n}{m}\Big(\mathbb{P}\left(X_1, \ldots, X_m \in M\backslash B_0, X_{m+1}, \ldots, X_n \in B_1\right) \\
&\qquad\qquad + \mathbb{P}\left(X_1, \ldots, X_m \in M\backslash B_0, X_{m+1}, \ldots, X_n \in B_2\right)\Big) \\
&= \sum_{m=0}^{n}\binom{n}{m}\left((1-P(B_0))^m P(B_1)^{n-m} + (1-P(B_0))^m P(B_2)^{n-m}\right) \\
&= (1-P(B_2))^n + (1-P(B_1))^n. \qquad \text{(C.26)}
\end{aligned}$$

Now we consider lower bounds for $P(B_1)$ and $P(B_2)$. Let $S_1 := \exp_{q_0}^{-1}(B_1) \cap \partial\mathcal{B}_{T_{q_0}M}(0, t_0)$, and as in Figure C.2, $\exp_{q_0}^{-1}(B_1) \subset \mathcal{B}_{T_{q_0}M}(0, t_0)$ is a cone satisfying

$$\frac{\mathcal{H}^d\left(\exp_{q_0}^{-1}(B_1)\right)}{\mathcal{H}^d\left(\mathcal{B}_{T_{q_0}M}(0,t_0)\right)} = \frac{\mathcal{H}^{d-1}(S_1)}{\mathcal{H}^{d-1}\left(\partial\mathcal{B}_{T_{q_0}M}(0,t_0)\right)}.$$

Let $\omega_d := \mathcal{H}^d(\mathcal{B}_{\mathbb{R}^d}(0,1))$ and $\sigma_d := \mathcal{H}^d(\partial\mathcal{B}_{\mathbb{R}^{d+1}}(0,1))$ be the volumes of the d-dimensional unit ball and the unit sphere respectively. Then by homogeneity, $\mathcal{H}^d\left(\mathcal{B}_{T_{q_0}M}(0,t_0)\right) = \omega_d t_0^d$ and $\mathcal{H}^{d-1}\left(\partial\mathcal{B}_{T_{q_0}M}(0,t_0)\right) = \sigma_{d-1}t_0^{d-1}$. For lower bounding $\mathcal{H}^{d-1}(S_1)$, let $v_0 := \frac{t_0\gamma_0'(0)}{\|\gamma_0'(0)\|} \in S_1$, and consider $\exp_{v_0} : T_{v_0}S_1 \to S_1$. Then $\tau_{S_1} = t_0$ and $\exp_{v_0}^{-1}(S_1) \subset \mathcal{B}_{T_{v_0}S_1}\left(0, \tau_{\min}^{-\frac{1}{2}}t_0^{\frac{3}{2}}\right)$, hence applying Proposition C.1 (v) yields

$$\begin{aligned}\mathcal{H}^{d-1}(S_1) &\geq \left(1 - \frac{t_0}{6\tau_{\min}}\right)^{d-1}\mathcal{H}^{d-1}\left(\mathcal{B}_{T_{v_0}S_1}\left(0, \tau_{\min}^{-\frac{1}{2}}t_0^{\frac{3}{2}}\right)\right)\\ &\geq \left(\frac{59}{60}\right)^{d-1}\omega_{d-1}\tau_{\min}^{-\frac{d-1}{2}}t_0^{\frac{3d-3}{2}},\end{aligned}$$

and hence

$$\begin{aligned}\mathcal{H}^{d-1}\left(\exp_{q_0}^{-1}(B_1)\right) &= \frac{\mathcal{H}^d\left(\mathcal{B}_{T_{q_0}M}(0,t_0)\right)\mathcal{H}^{d-1}(S_1)}{\mathcal{H}^{d-1}\left(\partial\mathcal{B}_{T_{q_0}M}(0,t_0)\right)}\\ &\geq \left(\frac{59}{60}\right)^{d-1}1\frac{\omega_{d-1}}{d}\tau_{\min}^{-\frac{d-1}{2}}t_0^{\frac{3d-1}{2}}.\end{aligned}$$

Furthermore, since $\exp_{q_0}^{-1}(B_1) \subset \mathcal{B}_{T_{q_0}M}(q_0, \frac{\tau_M}{10})$, Proposition C.1 (v) yields

$$\mathcal{H}^d(B_1) \geq \left(\frac{599}{600}\right)^d\mathcal{H}^d\left(\exp_{q_0}^{-1}(B_1)\right) \geq \left(\frac{35341}{36000}\right)^d\frac{1}{d}\tau_{\min}^{-\frac{d-1}{2}}t_0^{\frac{3d-1}{2}},$$

and hence,

$$P(B_1) \geq \left(\frac{35341}{36000}\right)^d\frac{f_{\min}}{d}\tau_{\min}^{-\frac{d-1}{2}}t_0^{\frac{3d-1}{2}} \geq C_{\tau_{\min},d,L,f_{\min}}s^{\frac{3d-1}{2}},$$

where $C_{\tau_{\min},d,L,f_{\min}} = \left(\frac{35341}{36000}\right)^d\frac{f_{\min}}{d}\tau_{\min}^{-\frac{d-1}{2}}\left(\frac{16\sqrt{2}}{(\sqrt{2}-1)\tau_{\min}^2} + \frac{(7\sqrt{2}-4)L}{3(\sqrt{2}-1)}\right)^{-1}$. By symmetry, the same bound holds for $P(B_2)$. Hence applying these to (C.26) gives

$$\begin{aligned}\mathbb{P}\left(\frac{1}{\tau_M} - \frac{1}{\hat{\tau}(X)} > s\right) &\leq 2\left(1 - C_{\tau_{\min},d,L,f_{\min}}s^{\frac{3d-1}{2}}\right)^n\\ &\leq 2\exp\left(-C_{\tau_{\min},d,L,f_{\min}}ns^{\frac{3d-1}{2}}\right).\end{aligned}$$

As a consequence, by integration,

$$\begin{aligned}\mathbb{E}_{P^n}\left[\left|\frac{1}{\hat{\tau}(X)} - \frac{1}{\tau_M}\right|^p\right] &= \int_0^{\frac{1}{\tau_M^p}}\mathbb{P}\left(\left|\frac{1}{\hat{\tau}(X)} - \frac{1}{\tau_M}\right|^p > s\right)ds\\ &\leq 2\int_0^\infty\exp\left(-C_{\tau_{\min},d,L,f_{\min}}ns^{\frac{3d-1}{2p}}\right)ds\\ &= 2\left(C_{\tau_{\min},d,L,f_{\min}}n\right)^{-\frac{2p}{3d-1}}\int_0^\infty x^{\frac{2p}{3d-1}}e^{-x}dx\\ &= 2\Gamma\left(\frac{2p}{3d-1}\right)C_{\tau_{\min},d,L,f_{\min}}^{-\frac{2p}{3d-1}}n^{-\frac{2p}{3d-1}}\\ &:= C_{\tau_{\min},d,L,f_{\min},p}n^{-\frac{2p}{3d-1}},\end{aligned}$$

where $\Gamma(\cdot)$ is the Gamma function. □

C.4 Minimax Lower Bounds

C.4.1 Stability of the Model With Respect to Diffeomorphisms

To prove Proposition V.30, we will use the following result (already reproduced in Lemma III.17), stating that the reach is a stable quantity with respect to $\mathcal{C}^2$-perturbations.

Lemma C.27 (Theorem 4.19 in [Fed59])**.** *Let $A \subset \mathbb{R}^D$ with $\tau_A \geq \tau_{min} > 0$ and $\Phi : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ is a $\mathcal{C}^1$-diffeomorphism such that Φ,Φ^{-1}, and $d\Phi$ are Lipschitz with Lipschitz constants K,N and R respectively, then*

$$\tau_{\Phi(A)} \geq \frac{\tau_{min}}{(K + R\tau_{min})N^2}.$$

Proof of Proposition V.30. Let $M' = \Phi(M)$ be the image of M by the mapping Φ. Since Φ is a global diffeomorphism, M' is a closed submanifold of dimension one. Moreover, Φ is $\|d\Phi\|_{op} \leq (1 + \|d\Phi - I_D\|_{op})$-Lipschitz, Φ^{-1} is $\left\|d\Phi^{-1}\right\|_{op} \leq (1 - \|d\Phi - I_D\|_{op})^{-1}$-Lipschitz, and $d\Phi$ is $\left\|d^2\Phi\right\|_{op}$-Lipschitz. From Lemma B.1,

$$\tau_{M'} \geq \frac{\tau_{min}(1 - \|d\Phi - I_D\|_{op})^2}{\left\|d^2\Phi\right\|_{op}\tau_{min} + (1 + \|d\Phi - I_D\|_{op})} \geq \tau_{min}/2,$$

where we used that $\left\|d^2\Phi\right\|_{op}\tau_{min} \leq 1/2$ and $\|d\Phi - I_D\|_{op} \leq 0.1$. All that remains to be proved now is the bound on the third order derivative of the geodesics of M'. We denote by γ and $\tilde{\gamma}$ the geodesics of M and M' respectively.

Let $p' = \Phi(p) \in M'$ and $v' = d_p\Phi.v \in T_{p'}M'$ be fixed. Since $M \in \mathcal{C}^{(3)}_{\tau_{min},L}$ is a compact $\mathcal{C}^3$-submanifold with geodesics $\|\gamma'''(0)\| \leq L$, M can be parametrized locally by a $\mathcal{C}^3$ bijective map $\Psi_p : \mathcal{B}_{\mathbb{R}^d}(0,\varepsilon) \to M$ with $\Psi_p(0) = p$. For a smooth curve γ on M nearby p, we let $c = (c_1, \ldots, c_d)^t$ denote its lift in the coordinates $\mathbf{x} = \Psi_p^{-1}$, that is $\gamma(t) = \Psi_p \circ c(t)$. $\gamma = \gamma_{p,v}$ is the geodesic of M with initial conditions p and v if and only if c satisfies the geodesic equations (see [dC92] p.62). That is, the second order ordinary differential equation

$$\begin{cases} c''_\ell(t) + \left\langle \Gamma^\ell\left(c(t)\right) \cdot c'(t), c'(t) \right\rangle = 0, & (1 \leq \ell \leq d) \\ c(0) = 0 \text{ and } c'(0) = d_p\mathbf{x}.v, \end{cases} \tag{C.28}$$

where $\Gamma^\ell = \left(\Gamma^\ell_{i,j}\right)_{1\leq i,j\leq d}$ are the Christoffel's symbols of the $\mathcal{C}^3$ chart $\mathbf{x}$, which depends only on $\mathbf{x}$ and its differentials of order 1 and 2. By construction, M' is parametrized locally by $\Psi'_{p'} = \Phi \circ \Psi_p$ yielding local coordinates $\mathbf{y} = \Psi'^{-1}_{p'} = \Psi_p^{-1} \circ \Phi^{-1}$ nearby $p' \in M'$. Writing $\tilde{\Gamma}^\ell$ for the Christoffel's symbols of M', $\tilde{\gamma}$ is a geodesic of M' at p' if its lift $\tilde{c} = \Psi'^{-1}_{p'}(\tilde{\gamma})$ satisfies (C.28) with Γ^ℓ replaced by $\tilde{\Gamma}^\ell$, and initial conditions $\tilde{c}(0) = c$ and $\tilde{c}'(0) = d_{p'}\mathbf{y}.v' = d_p\mathbf{x}.v$. From chain rule, the $\tilde{\Gamma}^\ell$'s depend on Γ, $d\Phi$, and $d^2\Phi$.

Considering $c'''(0) - \tilde{c}'''(0)$ by differentiating (C.28), since $c(0) = \tilde{c}(0) = 0$ and $c''(0) = \tilde{c}''(0)$, we have that for $\|I_D - d\Phi\|_{op}$, $\left\|d^2\Phi\right\|_{op}$ and $\left\|d^3\Phi\right\|_{op}$ small enough, this difference can be made arbitrarily small. In particular, $\tilde{\gamma}'''(0)$ is arbitrarily close to $\gamma'''(0)$ so that $\|\tilde{\gamma}'''(0)\| \leq \|\gamma'''(0)\| + L \leq 2L$, which concludes the proof. □

C.4.2 Some Lemmas on the Total Variation Distance

Prior to any actual construction, we show this straighforward lemma bounding the total variation between uniform distribution on manifolds that are perturbations of each other. For a d-submanifold $M \subset \mathbb{R}^D$, write $\lambda_M = \frac{\mathbb{1}_M}{\mathcal{H}^d(M)}\mathcal{H}^d$ for the uniform probability distribution on M.

Lemma C.29. *Let $M \subset \mathbb{R}^D$ be a d-dimensional submanifold and $\mathcal{B} \subset \mathbb{R}^D$ be a Borel set. Let $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ be a global diffeomorphism such that $\Phi_{|B^c}$ is the identity map and $\|d\Phi - I_D\|_{op} \leq 2^{1/d} - 1$. Then $\mathcal{H}^d(\Phi(M)) \leq 2\mathcal{H}^d(M)$ and $TV\big(\lambda_M, \lambda_{\Phi(M)}\big) \leq 12\lambda_M(B)$.*

Proof of Lemma C.29. Since Φ is $(1 + \|d\Phi - I_D\|_{op})$-Lipschitz, Lemma 4 of [ACLZ17] asserts that

$$\mathcal{H}^d\left(\Phi(M \cap B)\right) \leq (1 + \|d\Phi - I_D\|_{op})^d \mathcal{H}^d(M \cap B) \leq 2\mathcal{H}^d(M \cap B).$$

Therefore,

$$\begin{aligned}\mathcal{H}^d\left(\Phi(M)\right) - \mathcal{H}^d(M) &= \mathcal{H}^d\left(\Phi(M \cap B)\right) - \mathcal{H}^d\left(M \cap B\right) \\ &\leq \mathcal{H}^d(M \cap B) \leq \mathcal{H}^d(M).\end{aligned}$$

Now, writing $\triangle$ for the symmetric difference of sets, we have $M \triangle \Phi(M) = (B \cap M) \triangle (B \cap \Phi(M)) \subset (B \cap M) \cup (B \cap \Phi(M))$. Therefore, Lemma 5 in [ACLZ17] yields,

$$\begin{aligned}TV\big(\lambda_M, \lambda_{\Phi(M)}\big) &\leq 4\frac{\mathcal{H}^d\left(M \triangle \Phi(M)\right)}{\mathcal{H}^d(M \cup \Phi(M))} \\ &\leq 4\frac{\mathcal{H}^d\left(M \cap B\right) + \mathcal{H}^d\left(\Phi(M) \cap B\right)}{\mathcal{H}^d(M)} \\ &= 4\frac{\mathcal{H}^d\left(M \cap B\right) + \mathcal{H}^d\left(\Phi(M \cap B)\right)}{\mathcal{H}^d(M)} \\ &\leq 12\frac{\mathcal{H}^d(M \cap B)}{\mathcal{H}^d(M)} = 12\lambda_M(B).\end{aligned}$$

□

Let us now tackle the proof of Lemma V.29. For this, we will need the following elementary differential geometry results Lemma C.30 and Corollary C.31.

Lemma C.30. *Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be $\mathcal{C}^1$ and $x \in \mathbb{R}^d$ be such that $g(x) = 0$ and $d_x g \neq 0$. Then there exists $r > 0$ such that $\mathcal{H}^d\left(g^{-1}(0) \cap \mathcal{B}(x, r)\right) = 0$.*

Proof of Lemma C.30. Let us prove that for $r > 0$ small enough, the intersection $g^{-1}(0) \cap \mathcal{B}(x, r)$ is contained in a submanifold of codimension one of $\mathbb{R}^d$. Writing $g = (g_1, \ldots, g_k)$, assume without loss of generality that $\partial_{x_1} g_1 \neq 0$. Since $g_1 : \mathbb{R}^d \to \mathbb{R}$ is nonsingular at x, the implicit function theorem asserts that $g_1^{-1}(0)$ is a submanifold of dimension $d - 1$ of $\mathbb{R}^d$ in a neighborhood of $x \in \mathbb{R}^d$. Therefore, for $r > 0$ small enough, $g_1^{-1}(0) \cap \mathcal{B}(x, r)$ has d-Hausdorff measure zero. The result hence follows, noticing that $g^{-1}(0) \subset g_1^{-1}(0)$. □

Corollary C.31. *Let $M, M' \subset \mathbb{R}^D$ be two compact d-dimensional submanifolds, and $x \in M \cap M'$. If $T_x M \neq T_x M'$, there exists $r > 0$ such that $A = M \cap M' \cap \mathcal{B}(x, r)$ satisfies $\lambda_M(A) = \lambda_{M'}(A) = 0$.*

Proof of Corollary C.31. Writing $k = D - d$, we see that up to ambient diffeomorphism — which preserves the nullity of measure — we can assume that locally around x, M' coincides with $\mathbb{R}^d \times \{0\}^k$ and that M is the graph of a $\mathcal{C}^\infty$ function $g : \mathcal{B}_{\mathbb{R}^d}(0, r') \to \mathbb{R}^k$ for $r' > 0$ small enough. The assumption $T_x M \neq T_x M'$ translates to $d_0 g \neq 0$, and the previous transformation maps smoothly $M \cap M' \cap \mathcal{B}(x, r'')$ to $g^{-1}(0) \cap \mathcal{B}(0, r')$ for $r'' > 0$ small enough. We conclude by applying Lemma C.30. □

We are now in position to prove Lemma V.29.

Proof of Lemma V.29. Notice that P and P' are dominated by the measure $\mu = \mathbb{1}_{M\cup M'}\mathcal{H}^d$, with $dP(x) = f(x)d\mu(x)$ and $dP'(x) = f'(x)d\mu(x)$, where $f, f' : \mathbb{R}^D \to \mathbb{R}_+$ have support M and M' respectively. On the other hand, $\tilde{P}$ and $\tilde{P}'$ are dominated by $\nu(dx\,dT) = \delta_{\{T_xM, T_xM'\}}(dT)\,\mu(dx)$ with respective densities $\tilde{f}(x,T) = \mathbb{1}_{T=T_xM}f(x)$ and $\tilde{f}'(x,T) = \mathbb{1}_{T=T_xM'}f'(x)$, where we set arbitrarily $T_xM = T_0$ for $x \notin M$, and $T_xM' = T_0$ for $x \notin M'$. Recalling that f vanishes outside M, and f' outside M',

$$\begin{aligned} TV(\tilde{P}, \tilde{P}') &= \frac{1}{2}\int_{\mathbb{R}^D\times\mathbb{G}^{d,D}} |\bar{f} - \bar{f}'|d\nu \\ &= \frac{1}{2}\int_{\mathbb{R}^D} \mathbb{1}_{T_xM=T_xM'}|f(x) - f'(x)| + \mathbb{1}_{T_xM\neq T_xM'}(f(x) + f'(x))\mathcal{H}^d(dx). \end{aligned}$$

From Corollary C.31 and a straightforward compactness argument, we derive that

$$\mathcal{H}^d\left(M \cap M' \cap \{x | T_xM \neq T_xM'\}\right) = 0.$$

As a consequence, the above integral expression becomes

$$TV(\tilde{P}, \tilde{P}') = \frac{1}{2}\int_{\mathbb{R}^D} |f - f'|d\mathcal{H}^d = TV(P, P'),$$

which concludes the proof. □

C.4.3 Construction of the Hypotheses

This section is devoted to the construction of hypotheses that will be used in Le Cam's lemma (Lemma V.28), to derive Theorem V.13 and Theorem V.32.

Lemma C.32. *Let $R, \ell, \eta > 0$ be such that $\ell \leq \frac{R}{2} \wedge \left(2^{1/d} - 1\right)$ and $\eta \leq \frac{\ell^2}{2R}$. Then there exists a d-dimensional sphere of radius R that we call M, such that $M \in \mathcal{C}^{(3)}_{R,\frac{1}{R^2}}$ and a global $\mathcal{C}^\infty$-diffeomorphism $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ such that,*

$$\|d\Phi - I_2\|_{op} \leq \frac{3\eta}{\ell}, \quad \left\|d^2\Phi\right\|_{op} \leq \frac{23\eta}{\ell^2}, \quad \left\|d^3\Phi\right\|_{op} \leq \frac{573\eta}{\ell^3},$$

and so that writing $M' = \Phi(M)$, we have $\mathcal{H}^d(M') \leq 2\mathcal{H}^d(M) = 2\sigma_d R^d$

$$\left|\frac{1}{\tau_M} - \frac{1}{\tau_{M'}}\right| \geq \frac{\eta}{\ell^2}, \quad \text{and} \quad TV(\lambda_M, \lambda_{M'}) \leq 12\left(\frac{\ell}{R}\right)^d.$$

Proof of Lemma V.31. Let $M \subset \mathbb{R}^{d+1} \times \{0\}^{D-d-1} \subset \mathbb{R}^D$ be the sphere of radius R with center $(0, -R, 0, \ldots, 0)$. The reach of M is $\tau_M = R$, and its arc-length parametrized geodesics are arcs of great circles, which have third derivatives of constant norm $\|\gamma'''(t)\| = \frac{1}{R^2}$. Hence we see that $M \in \mathcal{C}^{(3)}_{R,\frac{1}{R^2}}$. Let ϕ be the map defined by $\phi(x) = \exp\left(\frac{\|x\|^2}{\|x\|^2-1}\right)\mathbb{1}_{\|x\|^2<1}$. ϕ is a symmetric $\mathcal{C}^\infty$ map with support equal to $\mathcal{B}(0,1)$ and elementary real analysis yields $\phi(0) = 1$, $\|d\phi\|_{op} \leq 3$, $\left\|d^2\phi\right\|_{op} \leq 23$ and $\left\|d^3\phi\right\|_{op} \leq 573$. Let $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ be defined by

$$\Phi(x) = x + \eta\phi\left(x/\ell\right) \cdot v,$$

where $v = (0, 1, 0, \ldots, 0)$ is the unit vertical vector. Φ is the identity map on $\mathcal{B}(0,\ell)^c$, and in $\mathcal{B}(0,\ell)$, Φ translates points on the vertical axis with a magnitude modulated by the weight function $\phi(x/\ell)$. From chain rule, $\|d\Phi - I_D\|_{op} = \eta\left\|d\phi\right\|_\infty/\ell \leq 3\eta/\ell < 1$. Therefore, $d_x\Phi$ is invertible for all $x \in \mathbb{R}^D$, so that Φ is a local $\mathcal{C}^\infty$-diffeomorphism according to the local

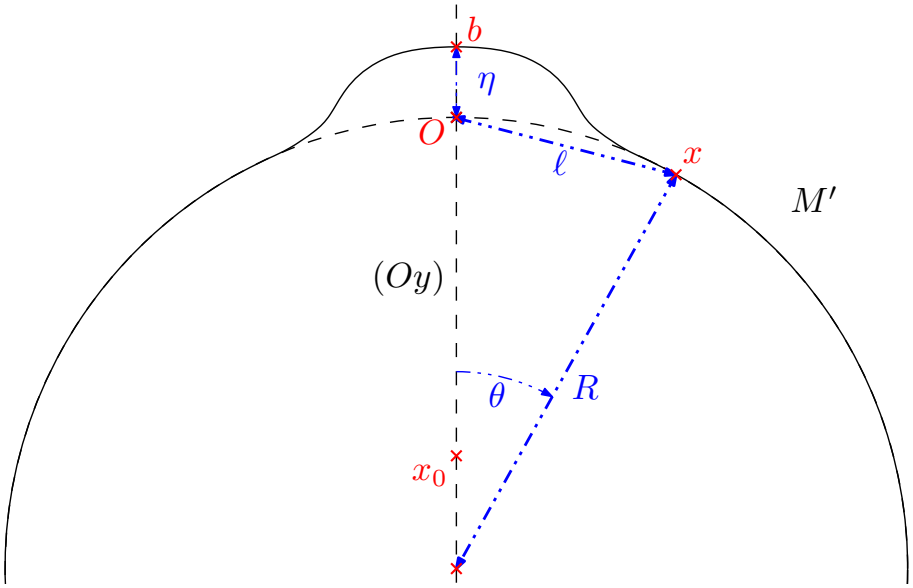


Figure C.3 – The bumped sphere circle M'.

inverse function theorem. Moreover, $\|\Phi(x)\| \to \infty$ as $\|x\| \to \infty$, so that Φ is a global $\mathcal{C}^\infty$-diffeomorphism by Hadamard-Cacciopoli theorem [DMGZ94]. Similarly, from bounds on differentials of ϕ we get

$$\left\|d^2\Phi\right\|_{op} \leq 23\frac{\eta}{\ell^2} \quad \text{and} \quad \left\|d^3\Phi\right\|_{op} \leq 573\frac{\eta}{\ell^3}.$$

Let us now write $M' = \Phi(M)$ for the image of M by the map Φ. Denote by (Oy) the vertical axis $\mathrm{span}(v)$, and notice that since ϕ is symmetric, M' is symmetric with respect to the vertical axis (Oy). We now bound from above the reach $\tau_{M'}$ of M' by showing that the point $x_0 = (0, (R+\eta/2)/(1+\frac{\ell^2}{2R\eta}), 0, \ldots, 0)$ belongs to its medial axis $Med(M')$. For this, write

$$b = (0, \eta, 0, \ldots, 0), \quad b' = (0, -2R, 0, \ldots, 0),$$

together with $\theta = \arccos(1 - \ell^2/(2R^2))$, and

$$x = (R\sin\theta, R\cos\theta - R, 0, \ldots, 0).$$

By construction, b, b' and x belong to M'. One easily checks that $\|x_0 - x\| < \|x_0 - b\|$ and $\|x_0 - x\| < \|x_0 - b'\|$, so that neither b nor b' is the nearest neighbor of x_0 on M'. But $x_0 \in (Oy)$ which is an axis of symmetry of M', and $(Oy) \cap M' = \{b, b'\}$. As a consequence, x_0 has strictly more than one nearest neighbor on M'. That is, x_0 belongs to the medial axis $Med(M')$ of M'. Therefore,

$$\begin{aligned}
\frac{1}{\tau_{M'}} &\geq \frac{1}{d(x_0, M')} \geq \frac{1}{\|x_0 - x\|} \\
&\geq \frac{1}{R\left|1 - \frac{\ell^2}{2R^2} - \frac{1+\frac{\eta}{2R}}{1+\frac{\ell^2}{2R\eta}}\right|} \\
&\geq \frac{1}{R\left(1 - \frac{1+\frac{\eta}{2R}}{1+\frac{\ell^2}{2R\eta}}\right)} \geq \frac{1}{R}\left(1 + \frac{1+\frac{\eta}{2R}}{1+\frac{\ell^2}{2R\eta}}\right) \geq \frac{1}{R} + \frac{\eta}{\ell^2},
\end{aligned}$$

which yields the bound $\left|\frac{1}{\tau_M} - \frac{1}{\tau_{M'}}\right| \geq \frac{\eta}{\ell^2}$.

Finally, since $M' = \Phi(M)$ with $\|d\Phi - I_D\|_{op} \leq 2^{1/d} - 1$ with $\Phi_{|\mathcal{B}(0,\ell)^c}$ coinciding with the identity map, Lemma C.29 yields $\mathcal{H}^d(M') \leq 2\mathcal{H}^d(M) = 2\sigma_d R^d$ and

$$\begin{aligned} TV(\lambda_M, \lambda_{M'}) &\leq 12\lambda_M(\mathcal{B}(0,\ell)) \\ &= 12\frac{\mathcal{H}^d\left(\mathcal{B}_{\mathcal{S}^d}\left(0, 2\arcsin\left(\frac{\ell}{2R}\right)\right)\right)}{\mathcal{H}^d(\mathcal{S}^d)} \\ &\leq 12\left(\frac{\ell}{R}\right)^d, \end{aligned}$$

which concludes the proof. □

Proof of Proposition V.31. Apply Lemma V.31 with $R = 2\tau_{min}$. Then the sphere M of radius $2\tau_{min}$ belongs to $\mathcal{C}^{(3)}_{2\tau_{min},1/(4\tau^2_{min})}$. Furthermore, taking $\eta = c_d\ell^3/\tau^2_{min}$ for $c_d > 0$ and $\ell > 0$ small enough, Proposition V.30 (applied to the unit sphere, yielding c_d, and reasoning by homogeneity for the sphere of radius $2\tau_{min}$) asserts that $M' = \Phi(M)$ belongs to $\mathcal{C}^{(3)}_{\tau_{min},1/(2\tau^2_{min})} \subset \mathcal{C}^{(3)}_{\tau_{min},L}$, since $L \geq 1/(2\tau^2_{min})$. Moreover,

$$\mathcal{H}^d(M')^{-1}, \mathcal{H}^d(M)^{-1} \geq (2^{d+1}\sigma_d\tau^d_{min}) \geq f_{min},$$

so that $\lambda_M, \lambda_{M'} \in \mathcal{P}^{(3)}_{\tau_{min},L}(f_{min})$, which gives the result. □

Let us now prove the minimax inconsistency of the reach estimation for $L = \infty$, using the same technique as above.

Proof of Proposition V.13. Let M and M' be given by Lemma C.32 with $\ell \leq \frac{R}{2} \wedge (2^{1/d} - 1)$, $\eta = \ell^2/(23R)$ and $R = 2\tau_{min}$. We have $\|d\Phi - I_D\|_{op} \leq 3\eta/\ell \leq 0.1$ and $\left\|d^2\Phi\right\|_{op} \leq 23\eta/\ell^2 \leq 1/(2\tau_{min})$. Since $\tau_M \geq 2\tau_{min}$, Lemma B.1 yields

$$\tau_{M'} \geq \frac{\tau_M(1 - \|d\Phi - I_D\|_{op})^2}{\left\|d^2\Phi\right\|_{op}\tau_M + (1 + \|d\Phi - I_D\|_{op})} \geq \tau_{min}.$$

As a consequence, M and M' belong to $\mathcal{C}^{(3)}_{\tau_{min},L=\infty}$. Furthermore, from $f_{min} \geq \frac{1}{2^{d+1}\tau^d_{min}\sigma_d} \geq \mathcal{H}^d(M)^{-1}, \mathcal{H}^d(M')^{-1}$ we see that the uniform distributions $\lambda_M, \lambda_{M'}$ belong to $\mathcal{P}^{(3)}_{\tau_{min},L=\infty}(f_{min})$. Let now $\tilde{P}, \tilde{P}'$ denote the distributions of $\tilde{\mathcal{P}}^{(3)}_{\tau_{min},L=\infty}(f_{min})$ associated to $\lambda_M, \lambda_{M'}$ (Definition V.9). Lemma V.29 asserts that $TV(\tilde{P}, \tilde{P}') = TV(\lambda_M, \lambda_{M'})$. Applying Lemma V.28 to $\tilde{P}, \tilde{P}'$, we get that for all $n \geq 1$, for ℓ small enough,

$$\begin{aligned} \inf_{\hat{\tau}_n} \sup_{\tilde{P} \in \tilde{\mathcal{P}}^{(3)}_{\tau_{min},L=\infty}(f_{min})} \mathbb{E}_{\tilde{P}^n}\left|\frac{1}{\tau_{\tilde{P}}} - \frac{1}{\hat{\tau}_n}\right|^p &\geq \frac{1}{2^p}\left|\frac{1}{\tau_M} - \frac{1}{\tau_{M'}}\right|^p (1 - TV(\lambda_M, \lambda_{M'}))^n \\ &\geq \frac{1}{2^p}\left(\frac{\eta}{\ell^2}\right)^p \left(1 - 12\left(\frac{\ell}{2\tau_{min}}\right)^d\right)^n \\ &= \frac{1}{2^p}\left(\frac{1}{46\tau_{min}}\right)^p \left(1 - 12\left(\frac{\ell}{2\tau_{min}}\right)^d\right)^n. \end{aligned}$$

Sending $\ell \to 0$ with $n \geq 1$ fixed yields the announced result. □

C.5 Stability with Respect to Tangent Spaces

Proof of Proposition V.34. To get the bound on the suprema, we show the (stronger) pointwise bound. For all $x, y \in \mathcal{X}$ with $x \neq y$,

$$\begin{aligned}\left|\frac{2d(y-x,T_x)}{\|y-x\|^2} - \frac{2d(y-x,T'_x)}{\|y-x\|^2}\right| &\leq \frac{2}{\|y-x\|^2}\|\pi_{T_x}(y-x) - \pi_{T'_x}(y-x)\| \\ &\leq \frac{2\|\pi_{T_x} - \pi_{T'_x}\|_{\mathrm{op}}}{\|y-x\|} \leq \frac{2\sin\theta}{\delta}.\end{aligned}$$

□

Chapter VI

Non-Asymptotic Rates for Manifold, Tangent Space and Curvature Estimation

Abstract

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

Content

VI.1 Introduction

In Chapter IV, we built manifold estimators achieving optimal rates of approximation over a class of $\mathcal{C}^2$ submanifolds. In the process, we studied tangent space estimators based on local PCA. This raises several new questions among which the optimality of the tangent space estimation procedure, as well as rates of convergence for smoother submanifolds. In addition, we tackled in Chapter V the estimation of the reach in a $\mathcal{C}^3$ model, yielding an estimation of maximum curvature in some cases. One can further ask about the estimation of curvature at all sample points, together with possibly better rates when the submanifold is smoother.

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

Among these three questions, a special attention has been paid to the estimation of the submanifold. In particular, it is a central problem in manifold learning. Indeed, there exists a wide bunch of algorithms intended to reconstruct submanifolds from point clouds (Isomap [TdSL00], LLE [RS00], and restricted Delaunay Complexes [BG14, CDR05] for instance), but a few come with theoretical guarantees [GPPVW12a]. Up to our knowledge, a minimax lower bound has proved optimality of a reconstruction scheme in only one case [GPPVW12a]. Some of these reconstruction procedures are based on tangent space estimation, such as in Chapter IV and [BG14, CDR05]. Tangent space estimation itself also yields interesting applications in manifold clustering [GM11, ACLZ17]. Estimation of curvature-related quantities naturally arises in shape reconstruction, since curvature can drive the size of a meshing. As a consequence, most of the associated results deal with the case $d = 2$ and $D = 3$, though some of them may be extended to higher dimensions [MOG11, GWM01]. Several algorithms have been proposed in that case [Rus04, CP05, MOG11, GWM01], but with no analysis of their performances from a statistical point of view.

To assess the quality of such a geometric estimator, the class of submanifolds over which the procedure is evaluated has to be specified. Up to now, the most commonly used model for submanifolds relied on the reach τ_M, a generalized convexity parameter. Assuming $\tau_M \geq \tau_{min} > 0$ involves both local regularity — a bound on curvature — and global regularity — no arbitrarily pinched area. This $\mathcal{C}^2$-like assumption has been extensively used in the computational geometry and geometric inference fields [NSW08, FLR+14, APR16, GPPVW12a]. One attempt of a specific investigation for higher orders of regularity $k \geq 3$ has been proposed in [CP05].

However, many works suggest that the regularity of the submanifold has an important impact on convergence rates. This is pretty clear for tangent space estimation, where convergence rates of PCA-based estimators range from $(1/n)^{1/d}$ in the $\mathcal{C}^2$ case (Chapter IV) to $(1/n)^{\alpha}$ with $1/d < \alpha < 2/d$ in more regular settings [SW12, TVF13]. In addition, it seems that PCA-based estimators are outperformed by estimators taking into account higher orders of smoothness [CC16, CP05], for regularities at least $\mathcal{C}^3$. For instance fitting quadratic terms lead to a convergence rate of order $(1/n)^{2/d}$ in [CC16]. These remarks naturally led us to investigate the properties of local polynomial approximation for regular submanifolds, where “regular" has to be properly defined. Local polynomial fitting for geometric inference was studied in several frameworks such as [CP05]. In some sense, a part of our work extends these results, by investigating the dependency of convergence rates on the sample size n, but also on the order of regularity k and the ambient and intrinsic dimensions d and D.

Outline

In this chapter, we build a collection of models for $\mathcal{C}^k$-submanifolds ($k \geq 3$) that naturally generalize the commonly used one for $k = 2$ (Section VI.2). We emphasize the necessity of both local and global constraints for estimation. On these models, we study the non-asymptotic rates of estimation for tangent space, second fundamental form, and submanifold estimation (Section VI.3). These results shed light on the influence of k, d, D and n on these estimation problems, showing for instance that the ambient dimension D plays no role. The estimators proposed all rely on the analysis of local polynomials, and allow to deal with the three estimation problems in a unified way (Section VI.4.1). Minimax lower bounds are derived using standard Bayesian techniques, although a new version of Assouad's Lemma is used for tangent spaces and second fundamental forms when the base point is random (Section VI.4.2). For the sake of completeness, geometric background and proofs of technical lemmas are given in the Appendix.

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

VI.2.1 Notation

Throughout this chapter, we consider d-dimensional compact submanifolds $M \subset \mathbb{R}^D$ without boundary. The submanifolds will always be assumed to be at least $\mathcal{C}^2$. For all $p \in M$, T_pM stands for the tangent space of M at p [dC92, Chapter 0]. We let $II_p^M : T_pM \times T_pM \to T_pM^\perp$ denote the second fundamental form of M at p [dC92, p. 125]. II_p^M characterizes the curvature of M at p. The standard inner product in $\mathbb{R}^D$ is denoted by $\langle \cdot, \cdot \rangle$ and the Euclidean distance by $\|\cdot\|$. Given a linear subspace $T \subset \mathbb{R}^D$, write $T^\perp$ for its orthogonal space. We write $\mathcal{B}(p, r)$ for the closed Euclidean ball of radius $r > 0$ centered at $p \in \mathbb{R}^D$, and for short $\mathcal{B}_T(p, r) = \mathcal{B}(p, r) \cap T$. For a smooth function $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ and $i \geq 1$, we let $d_x^i\Phi$ denote the ith order differential of Φ at $x \in \mathbb{R}^D$. For a linear map A defined on $T \subset \mathbb{R}^D$, $\|A\|_{\text{op}} = \sup_{v \in T} \frac{\|Av\|}{\|v\|}$ stands for the operator norm. We adopt the same notation $\|\cdot\|_{op}$ for tensors, i.e. multilinear maps. Similarly, if $\{A_x\}_{x \in T'}$ is a family of linear maps, for short, its L^∞ operator norm is denoted by $\|A\|_{op} = \sup_{x \in T'} \|A_x\|_{op}$. When it is well defined, we will write $\pi_B(z)$ for the projection of $z \in \mathbb{R}^D$ onto the closed subset $B \subset \mathbb{R}^D$, that is the nearest neighbor of z in B. The distance between two linear subspaces $U, V \subset \mathbb{R}^D$ of the same dimension is measured by the principal angle $\angle(U, V) = \|\pi_U - \pi_V\|_{\text{op}}$. The Hausdorff distance [GPPVW12a] in $\mathbb{R}^D$ is denoted by d_H. For a probability distribution P, $\mathbb{E}_P$ stands for the expectation with respect to P. We write $P^{\otimes n}$ for the n-times tensor product of P.

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

VI.2.2 Reach and Regularity of Submanifolds

As introduced in [Fed59], the reach τ_M of a subset $M \subset \mathbb{R}^D$ is the maximal neighborhood radius for which the projection π_M onto M is well defined. More precisely, denoting by $d(\cdot, M)$ the distance to M, the medial axis of M is defined to be the set of points which have at least two nearest neighbors on M, that is

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

The reach is then defined by

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

It gives a minimal scale of geometric and topological features of M. As a generalized convexity parameter, τ_M is a key parameter in reconstruction [GPPVW12a] and in topological inference [NSW08]. Having $\tau_M \geq \tau_{min} > 0$ prevents M from almost auto-intersecting, and bounds its curvature in the sense that $\left\| II_p^M \right\|_{op} \leq \tau_M^{-1} \leq \tau_{min}^{-1}$ for all $p \in M$ [NSW08, Proposition 6.1].

For $\tau_{min} > 0$, we let $\mathcal{C}^2_{\tau_{min}}$ denote the set of d-dimensional compact connected submanifolds M of $\mathbb{R}^D$ such that $\tau_M \geq \tau_{min} > 0$. A key property of submanifolds $M \in \mathcal{C}^2_{\tau_{min}}$ is the existence of a parametrization closely related to the projection onto tangent spaces. We let $\exp_p : T_pM \to M$ denote the geodesic map [dC92, Chapter 3], that is defined by $\exp_p(v) = \gamma_{p,v}(1)$, where $\gamma_{p,v}$ is the unique constant speed geodesic path of M with initial value p and velocity v.

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

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

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

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

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

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

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

In other words, elements of $\mathcal{C}^2_{\tau_{min}}$ have local parametrizations on top of their tangent spaces that are defined on neighborhoods with a minimal radius, and these parametrizations differ from the identity map by at most a quadratic term. In addition, the reach condition provides an order 2 Taylor expansion of the submanifold on top of its tangent spaces. A natural extension to $\mathcal{C}^k$-submanifolds should ensure that such an expansion exists at order k and satisfies some regularity constraints. To this aim, we introduce the following class of regularity $\mathcal{C}^k_{\tau_{min},\mathbf{L}}$.

Definition VI.2. *For $k \geq 3$, $\tau_{min} > 0$, and $\mathbf{L} = (L_\perp, L_3, \ldots, L_k)$, we let $\mathcal{C}^k_{\tau_{min},\mathbf{L}}$ denote the set of d-dimensional compact connected submanifolds M of $\mathbb{R}^D$ with $\tau_M \geq \tau_{min}$ and such that, for all $p \in M$, there exists a local one-to-one parametrization Ψ_p of the form:*

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

for some $r \geq \frac{1}{8L_\perp}$, with $\mathbf{N}_p \in \mathcal{C}^k\left(\mathcal{B}_{T_pM}\left(0, r\right), \mathbb{R}^D\right)$ such that

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

for all $\|v\| \leq \frac{1}{8L_\perp}$. Furthermore, we require that

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

Let us precise that the radius $1/(8L_\perp)$ has been chosen for convenience. Other smaller scales would do and we could even parametrize this constant, but without substantial benefits in the results.

The Ψ_p's can be seen as unit parametrizations of M. The conditions on $\mathbf{N}_p(0)$, $d_0\mathbf{N}_p$, and $d_v^2\mathbf{N}_p$ ensure that Ψ_p^{-1} is close to the projection π_{T_pM}. The bounds on $d_v^i\mathbf{N}_p$ $(3 \le i \le k)$ allow to control the coefficients of the polynomial expansion we seek. Indeed, whenever $M \in \mathcal{C}^k_{\tau_{min},\mathbf{L}}$, Lemma VI.14 shows that for every p in M, and y in $\mathcal{B}(p, \frac{\tau_{min}\wedge L_\perp^{-1}}{4}) \cap M$,

$$y - p = \pi^*(y-p) + \sum_{i=2}^{k-1} T_i^*(\pi^*(y-p)^{\otimes i}) + R_k(y-p), \tag{VI.3}$$

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

Such Ψ_p's exist for any compact $\mathcal{C}^k$-submanifold, if one allows τ_{min}^{-1}, $L_\perp$, L_3,...,L_k to be large enough. Note that for $k \ge 3$ the exponential map can happen to be only $\mathcal{C}^{k-2}$ for a $\mathcal{C}^k$-submanifold [Har51]. Hence, it may not be a good choice of Ψ_p. However, for $k = 2$, taking $\Psi_p = \exp_p$ is sufficient for our purpose. For ease of notation, we may write $\mathcal{C}^2_{\tau_{min},\mathbf{L}}$ although the specification of $\mathbf{L}$ is useless. In this case, we implicitly set by default $\Psi_p = \exp_p$ and $L_\perp = 5/(4\tau_{min})$.

As will be shown in Theorem VI.6, the global assumption $\tau_M \ge \tau_{min} > 0$ cannot be dropped, even when higher order regularity bounds L_i's are fixed.

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

Definition VI.4. *For $k \ge 2$, $\tau_{min} > 0$, $\mathbf{L} = (L_\perp, L_3, \ldots, L_k)$ and $f_{min} \le f_{max}$, we let $\mathcal{P}^k_{\tau_{min},\mathbf{L}}(f_{min}, f_{max})$ denote the set of distributions P with support $M \in \mathcal{C}^k_{\tau_{min},\mathbf{L}}$ that have a density f with respect to the volume measure on M, and such that for all $x \in M$,*

$$0 < f_{min} \le f(x) \le f_{max} < \infty.$$

For short, we write $\mathcal{P}^k$ when there is no ambiguity. We denote by $\mathbb{X}_n$ an i.i.d. n-sample $\{X_1, \ldots, X_n\}$, that is, a sample with distribution $P^{\otimes n}$ for some $P \in \mathcal{P}^k$. In what follows, though M is unknown, all the parameters of the model will be assumed to be known, including the intrinsic dimension d and the order of regularity k. We will also denote by $\mathcal{P}^k_{(x)}$ the subset of elements in $\mathcal{P}^k$ whose support contains a prescribed $x \in \mathbb{R}^D$.

In view of our minimax study on $\mathcal{P}^k$, it is important to ensure by now that $\mathcal{P}^k$ is stable with respect to deformations and dilations. Here, since we deal with submanifolds, a natural way to model deformations is through ambient diffeomorphisms.

Proposition VI.5. *Let $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ be a global $\mathcal{C}^k$-diffeomorphism. If $\|d\Phi - I_D\|_{op}$, $\|d^2\Phi\|_{op}$, ..., $\left\|d^k\Phi\right\|_{op}$ are small enough, then for all P in $\mathcal{P}^k_{\tau_{min},\mathbf{L}}(f_{min}, f_{max})$, the pushforward distribution $P' = \Phi_* P$ belongs to $\mathcal{P}^k_{\tau_{min}/2,2\mathbf{L}}(f_{min}/2, 2f_{max})$. Moreover, if $\Phi = \lambda I_D$ $(\lambda > 0)$ is an homogeneous dilation, then $P' \in \mathcal{P}^k_{\lambda\tau_{min},\mathbf{L}_{(\lambda)}}(f_{min}/\lambda^d, f_{max}/\lambda^d)$, where $\mathbf{L}_{(\lambda)} = (L_\perp/\lambda, L_3/\lambda^2, \ldots, L_k/\lambda^{k-1})$.*

Proposition VI.5 follows from a geometric reparametrization argument (Proposition D.2) and a change of variable result for the Hausdorff measure (Lemma D.3).

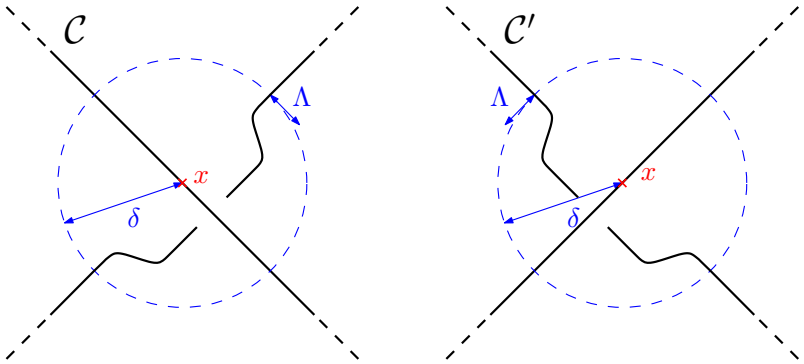


Figure VI.1 – Inconsistency of tangent space estimation for $\tau_{min} = 0$.

VI.2.3 Necessity of a Global Assumption

In the previous Section VI.2.2, we generalized $\mathcal{C}^2$-like models — stated in terms of reach — to $\mathcal{C}^k$ for $k \geq 3$ by imposing higher order differentiability bounds on parametrizations Ψ_p's. Though, we did not drop the global assumption $\tau_M \geq \tau_{min} > 0$. Indeed, it appears that such an assumption is necessary for estimation purpose.

Theorem VI.6. *Assume that $\tau_{min} = 0$. If $D \geq d + 3$, then for all $k \geq 3$ and $L_\perp > 0$, provided that $L_3/L_\perp^2, \ldots, L_k/L_\perp^{k-1}, L_\perp^d/f_{min}$ and $f_{max}/L_\perp^d$ are large enough (depending only on d and k), for all $n \geq 1$,*

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

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

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

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

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

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

VI.3 Main Results

Let us now move to the description of the main results, that consist of minimax upper and lower bounds for each object of interest. Given an i.i.d. n-sample $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ with unknown common distribution $P \in \mathcal{P}^k$ having support M, we detail non-asymptotic rates for the estimation of tangent spaces $T_{X_j} M$, second fundamental forms $II_{X_j}^M$, and M itself.

For this, we need one more piece of notation. For $1 \leq j \leq n$, $P_n^{(j)}$ denotes integration with respect to $1/(n-1) \sum_{i \neq j} \delta_{(X_i - X_j)}$, and $y^{\otimes i}$ denotes the $D \times i$-dimensional vector

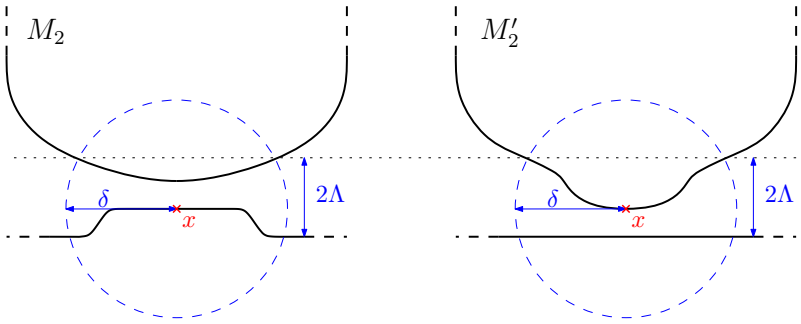


Figure VI.2 – Inconsistency of curvature estimation for $\tau_{min} = 0$.

$(y, \dots, y)$. For a constant $t > 0$ and a bandwidth $h > 0$ to be chosen later, we define the local polynomial estimator $(\hat{\pi}_j, \hat{T}_{2,j}, \dots, \hat{T}_{k-1,j})$ at X_j to be any element of

$$\underset{\pi, \sup_{2\le i\le k}\|T_i\|_{op}\le t}{\arg\min} P_n^{(j)} \left[\left\| x - \pi(x) - \sum_{i=2}^{k-1} T_i(\pi(x)^{\otimes i}) \right\|^2 \mathbb{1}_{\mathcal{B}(0,h)}(x) \right], \tag{VI.7}$$

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

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

Note that the set of d-dimensional orthogonal projectors is not convex, leading to a more involved optimization problem than usual least squares. In practice, this problem may be solved using tools from optimization on Grassman manifolds [UM14], or adopting a two-stage procedure such as in [CP05]: from local PCA, a first d-dimensional space is estimated at each sample point, along with an orthonormal basis of it. Then, the optimization problem (VI.7) is expressed as a minimization problem in terms of the coefficients of $(\pi_j, T_{2,j}, \dots, T_{k,j})$ in this basis under orthogonality constraints. It is worth mentioning that a similar problem is explicitly solved in [CC16], leading to an optimal tangent space estimation procedure in the case $k = 3$.

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

VI.3.1 Tangent Spaces

By definition, the tangent space $T_{X_j}M$ is the best linear approximation of M nearby X_j. Therefore, it is natural to take the range of the first order term minimizing (VI.7) and write $\hat{T}_j = \operatorname{im} \hat{\pi}_j$. The $\hat{T}_j$'s approximate simultaneously the $T_{X_j}M$'s with high probability, as stated below.

Theorem VI.8. *Assume that* $t \geq C_{k,d,\tau_{min},\mathbf{L}} \geq \sup_{2\leq i\leq k} \|T_i^*\|_{op}$. *Set* $h = \left(C_{d,k}\frac{\log(n) f_{max}^2}{(n-1)f_{min}^3}\right)^{1/d}$, *for* $C_{d,k}$ *large enough. If* n *is large enough so that* $h \leq h_0$, *then with probability at least* $1 - \left(\frac{1}{n}\right)^{k/d}$,

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

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

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

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

The same bound holds for the estimation of T_xM at a prescribed $x \in M$. For that, simply take $P_n^{(x)} = 1/n \sum_i \delta_{(X_i - x)}$ as integration in (VI.7).

This result is in line with those of [CP05] in terms of the sample size dependency $(1/n)^{(k-1)/d}$. Besides, it shows that the convergence rate of our estimator does not depend on the ambient dimension D, even in codimension greater than 2. When $k = 2$, we recover the same rate as in Chapter IV, where we used local PCA for estimation, that is a reformulation of (VI.7). When $k \geq 3$, this procedure outperforms PCA-based estimators of [SW12] and [TVF13], where convergence rates of the form $(1/n)^\alpha$ is obtained for $1/d < \alpha < 2/d$. This bound also recovers the result of [CC16] in the case $k = 3$, where a similar procedure is used. Moreover, Theorem VI.8 nearly matches the following lower bound.

Theorem VI.9. *If* $\tau_{min}L_\perp, \ldots, \tau_{min}^{k-1}L_k, (\tau_{min}^d f_{min})^{-1}$ *and* $\tau_{min}^d f_{max}$ *are large enough (depending only on* d *and* k*), then*

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

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

Hence, up to a $\log n$ factor, the rate $n^{-(k-1)/d}$ is optimal for tangent space estimation on the model $\mathcal{P}^k$. The rate $(\log n/n)^{1/d}$ obtained in Chapter IV for $k = 2$ is therefore optimal, as well as the rate $(\log n/n)^{2/d}$ given in [CC16] for $k = 3$. The rate $n^{-(k-1)/d}$ naturally appears on the the model $\mathcal{P}^k$, since it consists of $\mathcal{C}^k$-submanifolds, and tangent spaces are differential objects of order 1, yielding the shift $k - 1$. Again, the same lower bound holds for the estimation of T_xM at a fixed point x in the model $\mathcal{P}_{(x)}^k$. Interestingly, the tools used to derive the lower bound for T_xM (x fixed) is much less involved than for $T_{X_1}M$ (X_1 random and depending on the distribution P). In the latter case, a conditional Assouad's Lemma (Lemma VI.24) is used. We will detail these differences in Section VI.4.2.

VI.3.2 Curvature

The second fundamental form $II_{X_j}^M : T_{X_j}M \times T_{X_j}M \to T_{X_j}M^\perp \subset \mathbb{R}^D$ is a symmetric bilinear map that encodes completely the curvature of M at X_j [dC92, Chap. 6, Proposition 3.1]. Estimating it only from a point cloud $\mathbb{X}_n$ does not trivially make sense, since $II_{X_j}^M$

has domain $T_{X_j}M$ which is unknown. To bypass this issue we extend $II^M_{X_j}$ to $\mathbb{R}^D$. That is, we consider the estimation of $II^M_{X_j} \circ \pi_{T_{X_j}M}$ which has full domain $\mathbb{R}^D$. Following the same ideas as in the previous Section VI.3.1, we use the second order tensor $\hat{T}_{2,j} \circ \hat{\pi}_j$ obtained in (VI.7) to estimate $II^M_{X_j} \circ \pi_{T_{X_j}M}$.

Theorem VI.10. *Let $k \geq 3$. Take h as in Theorem VI.8 and $t = 1/h$. If n is large enough so that $h \leq h_0$ and $h^{-1} \geq C^{-1}_{k,d,\tau_{min},\mathbf{L}} \geq (\sup_{2\leq i\leq k} \|T_i^*\|_{op})^{-1}$, then with probability at least $1 - \left(\frac{1}{n}\right)^{k/d}$,*

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

In particular, for n large enough,

$$\sup_{P\in\mathcal{P}^k} \mathbb{E}_{P^{\otimes n}} \max_{1\leq j\leq n} \left\| II^M_{X_j} \circ \pi_{T_{X_j}M} - \hat{T}_{2,j} \circ \hat{\pi}_j \right\|_{op} \leq C \left(\frac{\log(n)}{n-1}\right)^{\frac{k-2}{d}},$$

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

Interestingly, Theorems VI.8 and VI.10 are enough to provide estimators of various notions of curvature. For instance, consider the scalar curvature [dC92, Section 4.4] at a point X_j, defined by

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

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

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

where $(\hat{e}_r)_{1\leq r\leq d}$ is an orthonormal basis of $\hat{T}_{X_j}M$. Theorems VI.8 and VI.10 yield

$$\mathbb{E}_{P^{\otimes n}} \max_{1\leq j\leq n} \left| \widehat{Sc}_j - Sc^M_{X_j} \right| \leq C_{d,k,\tau_{min},\mathbf{L},f_{min},f_{max}} \left(\frac{\log(n)}{n-1}\right)^{\frac{k-2}{d}}.$$

The (near-)optimality of the bound stated in Theorem VI.10 is assessed by the following lower bound.

Theorem VI.11. *If $\tau_{min}L_\perp, \ldots, \tau^{k-1}_{min}L_k, (\tau^d_{min}f_{min})^{-1}$ and $\tau^d_{min}f_{max}$ are large enough (depending only on d and k), then*

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

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

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

Notice that the lower bound (Theorem VI.11) does not require $k \geq 3$. Hence, we get that for $k = 2$, curvature cannot be estimated uniformly consistently on the $\mathcal{C}^2$-model $\mathcal{P}^2$. This seems natural, since the estimation of a second order quantity should require an additional degree of smoothness.

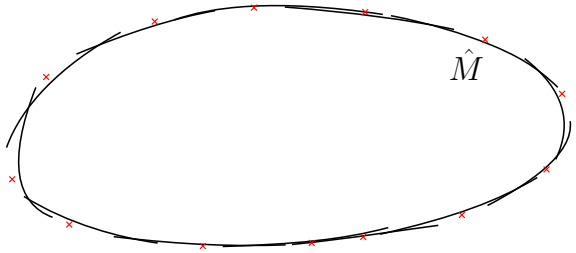


Figure VI.3 – $\hat{M}_{\texttt{POLY}}$ is a union of polynomial patches at sample points.

VI.3.3 Support Estimation

For each $1 \leq j \leq n$, the minimization (VI.7) outputs a series of tensors $(\hat{\pi}_j, \hat{T}_{2,j}, \ldots, \hat{T}_{k-1,j})$. This collection of multidimensional monomials can be further exploited as follows. By construction, they fit M at scale h around X_j, so that

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

is a good candidate for an approximate parametrization in a neighborhood of X_j. We do not know the domain $T_{X_j}M$ of the initial parametrization, though we have at hand an approximation $\hat{T}_j = \operatorname{im}\hat{\pi}_j$ which was proved to be consistent in Section VI.3.1. As a consequence, we let the support estimator based on local polynomials $\hat{M}_{\texttt{POLY}}$ be

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

The set $\hat{M}_{\texttt{POLY}}$ has no reason to be globally smooth, since it consists of a union of polynomial patches that are not linked together (Figure VI.3). However, $\hat{M}_{\texttt{POLY}}$ is provably close to M for the Hausdorff distance.

Theorem VI.12. *With the same assumptions as Theorem VI.10, with probability at least* $1 - 2\left(\frac{1}{n}\right)^{\frac{k}{d}}$*, we have*

$$d_H(M, \hat{M}_{\texttt{POLY}}) \leq C_{d,k,\tau_{min},\mathbf{L},f_{min},f_{max}} h^k.$$

In particular, for n *large enough,*

$$\sup_{P \in \mathcal{P}^k} \mathbb{E}_{P^{\otimes n}} d_H(M, \hat{M}_{\texttt{POLY}}) \leq C\left(\frac{\log(n)}{n-1}\right)^{\frac{k}{d}},$$

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

For $k = 2$, we recover the rate $(\log n/n)^{2/d}$ obtained in [GPPVW12a, KZ15] and Chapter IV. However, our estimator $\hat{M}_{\texttt{POLY}}$ is an unstructured union of d-dimensional balls in $\mathbb{R}^D$. Consequently, $\hat{M}_{\texttt{POLY}}$ does not recover the topology of M as the estimators $\hat{M}_{\texttt{TDC}}$, $\hat{M}_{\texttt{TDC}\delta}$ and $\hat{M}_{\texttt{TDC}}+$ of IV do.

When $k \geq 3$, $\hat{M}_{\texttt{POLY}}$ outperforms reconstruction procedures based on a somewhat piecewise linear interpolation (for instance in [GPPVW12a] and Chapter IV), and achieves the faster rate $(\log n/n)^{k/d}$ for the Hausdorff loss. This seems quite natural, since our procedure fits higher order terms. This is done at the price of a probably worse dependency on the dimension d than in [GPPVW12a] and Chapter IV. Theorem VI.12 is now proved to be (almost) minimax optimal.

Theorem VI.13. *If $\tau_{min}L_\perp, \dots, \tau_{min}^{k-1}L_k, (\tau_{min}^d f_{min})^{-1}$ and $\tau_{min}^d f_{max}$ are large enough (depending only on d and k), then for n large enough,*

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

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

Theorem VI.13 is obtained from Le Cam's Lemma (Theorem VI.22). Let us note that it is likely for the extra $\log n$ term appearing in Theorem VI.12 to actually be present in the minimax rate. Roughly, it is due to the fact that the Hausdorff distance d_H is similar to a L^∞ loss. The $\log n$ term may be obtained in Theorem VI.13 with the same combinatorial analysis as in [KZ15] for $k = 2$.

VI.4 Main Ideas of the Proofs

VI.4.1 Local Polynomials

We now turn to the proof of the upper bounds of Section VI.3. First, to relate the existence of parametrizations Ψ_p's to a local polynomial decomposition, the following lemma is needed.

Lemma VI.14. *For any $M \in \mathcal{C}^k_{\tau_{min},\mathbf{L}}$ and $x \in M$, the following holds.*

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

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

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

$$M \cap \mathcal{B}\left(x, \frac{3h}{5}\right) \subset \Psi_x\left(\mathcal{B}_{T_xM}(x, h)\right) \subset M \cap \mathcal{B}\left(x, \frac{5h}{4}\right).$$

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

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

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

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

with

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

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

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

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

The proof of Lemma VI.14 can be found in Section D.1.2. We are now in position to analyze local polynomial estimators. For clarity's sake, the bounds are given for $j = 1$, where we denote by $\hat{\pi}$, $\hat{T}_i$ $(2 \leq i \leq k-1)$ the fitted polynomials of (VI.7), and $P_{n-1} = P_{n-1}^{(1)}$. The results of Theorems VI.8, VI.10, and VI.12 then follow from a straightforward union bound. We also set $k \geq 3$, the case $k = 2$ proceeding from the same derivation, omitting the higher order tensors. Without loss of generality, we can assume that $X_1 = 0$ and that T_0M is spanned by the first d vectors of the canonical basis, so that $\pi^*(x) = (x_1, \ldots, x_d, 0, \ldots, 0) = (x_{1:d}, 0, \ldots, 0)$.

Recall that $h_0 = (\tau_{min} \wedge L_\perp^{-1})/8$. According to Lemma VI.14, if $M \in \mathcal{C}^k_{\tau_{min},\mathbf{L}}$, for any $x \in M$ such that $\|x\| \leq h_0$, we may write

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

where $\|R_k(x)\| \leq C_{\tau_{min},\mathbf{L}}\|x\|^k$. Every coordinate of $(\hat{T}_i - T_i^*)(\pi^*(x))$ may be thought of as a polynomial map in the variable $x_{1:d}$. Thus, proximity between $\hat{T}_i$ and T_i^* will be first stated in terms of polynomial norm.

Let $\mathbb{R}^k[x_{1:d}]$ denote the set of real-valued polynomial functions in d variables with degree less than k. For $Q \in \mathbb{R}^k[x_{1:d}]$, we denote by $\|Q\|_2$ the Euclidean norm of its coefficients, and by Q_h the polynomial defined by $Q_h(x_{1:d}) = Q(hx_{1:d})$. The following result relates the $L^2(P_{n-1})$ norm involved in (VI.7) to polynomial norms.

Proposition VI.15. *Set* $h = \left(K\frac{\log(n)}{n-1}\right)^{\frac{1}{d}}$. *There exist constants* $\kappa_{k,d}$, $c_{k,d}$ *and* C_d *such that, if* $K \geq (\kappa_{k,d} f_{max}^2/f_{min}^3)$ *and* n *is large enough so that* $h \leq h_0 \leq \tau_{min}/4$, *then with probability at least* $1 - \left(\frac{1}{n}\right)^{\frac{k}{d}+1}$, *we have*

$$\begin{aligned} P_{n-1}[Q^2(\pi^*(x))\mathbb{1}_{\mathcal{B}(h)}(x)] &\geq c_{k,d} h^d f_{min} \|Q_h\|_2^2, \\ N(h) &\leq C_d f_{max}(n-1)h^d, \end{aligned}$$

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

The proof of Proposition VI.15 is deferred to Section D.2.2. From now on we assume that the probability event defined in Proposition VI.15 occurs. For short, with a slight abuse of notation, we denote by $T_{p:q}(x)$ the sum $T_p(x^{\otimes p}) + \ldots + T_q(x^{\otimes q})$, and by $\mathcal{R}_{n-1}(\pi, T_2, \ldots, T_{k-1})$ the empirical criterion defined by (VI.7). Since for $t \geq \max_{i=2,\ldots,k-1} \|T_i^*\|_{op}$,

$$\mathcal{R}_{n-1}(\hat{\pi}, \hat{T}_1, \ldots, \hat{T}_{k-1}) \leq \mathcal{R}_{n-1}(\pi^*, T_2^*, \ldots, T_{k-1}^*) \leq C_{\tau_{min},\mathbf{L}} h^{2k} N(h)/(n-1)$$

according to (VI.3), we may write

$$\begin{aligned} C_{\tau_{min},\mathbf{L}} h^{2k} \frac{N(h)}{n-1} &\geq \mathcal{R}_{n-1}(\hat{\pi}, \hat{T}_2, \ldots, \hat{T}_{k-1}) \\ &= P_{n-1}\left(\left\|(\pi^* - \hat{\pi})(x) + (T_{2:k-1}^* \circ \pi^* - \hat{T}_{2:k-1} \circ \hat{\pi})(x) \right.\right. \\ &\qquad \left.\left. + R_k(x)\right\|^2 \mathbb{1}_{\mathcal{B}(0,h)}(x)\right), \end{aligned}$$

with $\|R_k(x)\| \leq C_{\tau_{min},\mathbf{L}} h^{2k}$. It follows that

$$P_{n-1}\left(\left\|(\pi^* - \hat{\pi})(x) + (T^*_{2:k-1} \circ \pi^* - \hat{T}_{2:k-1} \circ \hat{\pi})(x)\right\|^2 \mathbb{1}_{\mathcal{B}(0,h)}(x)\right) \leq C_{\tau_{min},\mathbf{L}} h^{2k} \frac{N(h)}{n-1} \leq C_{\tau_{min},\mathbf{L},d} f_{max} h^{d+2k}.$$

On the other hand, using (VI.3) again yields, for $x \in \mathcal{B}(0,h) \cap M$,

$$(\pi^* - \hat{\pi})(x) + (T^*_{2:k-1} \circ \pi^* - \hat{T}_{2:k-1} \circ \hat{\pi})(x) = T'_1(\pi^*(x)) + T'_2(\pi^*(x)^{\otimes 2}) + T'_{3:k}(\pi^*(x)) + \hat{\pi}(R_k(x)) + R'_k(x),$$

with $\|R_k(x)\| \leq C_{\tau_{min},\mathbf{L}} h^k$, $\|R'_k(x)\| \leq t C_{\tau_{min},k,\mathbf{L}} h^{k+1}$ since only tensors of order greater than 2 are involved in R'_k, and

$$\begin{array}{rcl} T'_1(\pi^*(x)) & = & (\pi^* - \hat{\pi})\pi^*(x) \\ T'_2(\pi^*(x)^{\otimes 2}) & = & (\pi^* - \hat{\pi})(T^*_2(\pi^*(x)^{\otimes 2})) + (T^*_2 \circ \pi^* - \hat{T}_2 \circ \hat{\pi})(\pi^*(x)^{\otimes 2}). \end{array}$$

Hence,

$$P_{n-1}\left(\left\|T'_1(\pi^*(x)) + T'_2(\pi^*(x)^{\otimes 2}) + T'_{3:k}(\pi^*(x))\right\|^2 \mathbb{1}_{\mathcal{B}(0,h)}(x)\right) \leq C_{\tau_{min},\mathbf{L},d} f_{max} h^{d+2k} (1 + ht). \quad \text{(VI.16)}$$

The left-hand side of (VI.16) may be decomposed coordinate-wise as

$$P_{n-1}\left(\left\|T'_1(\pi^*(x)) + T'_2(\pi^*(x)^{\otimes 2}) + T'_{3:k}(\pi^*(x))\right\|^2 \mathbb{1}_{\mathcal{B}(0,h)}(x)\right) = \sum_{j=1}^{D} P_{n-1}\left(\left(T'^{(j)}_1(\pi^* x) + T'^{(j)}_2(\pi^*(x)^{\otimes 2}) + T'^{(j)}_{3:k}(\pi^*(x))\right)^2 \mathbb{1}_{\mathcal{B}(0,h)}(x)\right),$$

where for any tensor T, $T^{(j)}$ denotes the j-th coordinate of T and is considered as a real valued j-order polynomial. Then, for every j, Proposition VI.15 leads to

$$\begin{aligned} P_{n-1}\left(\left(T'^{(j)}_1(\pi^*(x)) + T'^{(j)}_2(\pi^*(x)^{\otimes 2}) + T'^{(j)}_{3:k}(\pi^*(x))\right)^2 \mathbb{1}_{\mathcal{B}(0,h)}(x)\right) &\geq c_{d,k} f_{min} h^d \left\|\left(T'^{(j)}_1(\pi^*(x)) + T'^{(j)}_2(\pi^*(x)^{\otimes 2}) + T'^{(j)}_{3:k}(\pi^*(x))\right)_h\right\|_2^2 \\ &= c_{d,k} f_{min} h^d \sum_{i=1}^{k} \left\|\left(T'^{(j)}_i(\pi^*(x)^{\otimes i})\right)_h\right\|_2^2. \end{aligned}$$

Summing all contributions leads to

$$c_{d,k} f_{min} \sum_{j=1}^{D} \sum_{i=1}^{k} \left\|\left(T'^{(j)}_i(\pi^*(x)^{\otimes i})\right)_h\right\|_2^2 \leq C_{k,\mathbf{L},d,\tau_{min}} f_{max} h^{2k}(1 + t^2 h^2).$$

This entails

$$\|T'_i\|_{op}^2 \leq C_{d,k,\mathbf{L},\tau_{min}} \frac{f_{max}}{f_{min}} h^{2(k-i)}(1 + t^2 h^2), \quad \text{(VI.17)}$$

for $1 \leq i \leq k$, as well as

$$\left\|(\pi^* - \hat{\pi})(x) + (T^*_{2:k-1} \circ \pi^* - \hat{T}_{2:k-1} \circ \hat{\pi})(x)\right\| \leq C_{d,k,\mathbf{L},\tau_{min}} \sqrt{\frac{f_{max}}{f_{min}}} h^k (1 + th), \quad \text{(VI.18)}$$

for $x \in \mathcal{B}(0,h) \cap M$, according to (VI.3).

Bounds for Tangent Space Estimation

Noting that

$$\|T_1'\|_{op} = \|(\pi^* - \hat{\pi})\pi^*\|_{op} = \|\pi_{\hat{T}_1^\perp} \circ \pi^*\| = \angle(T_0M, \hat{T}_1)$$

from Proposition III.29, and using (VI.17) for $i = 1$ yields Theorem VI.8.

Bounds for Curvature Estimation

In accordance with assumptions of Theorem VI.10, we assume that $\max_{2\leq i\leq k} \|T_i^*\|_{op} \leq t \leq 1/h$. Since

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

we deduce that

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

Using (VI.17) with $i = 1, 2$ and $th \leq 1$ leads to

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

Finally, Lemma VI.14 states that $II_{X_1}^M = T_2^*$, hence Theorem VI.10 is proved.

Bounds for Reconstruction

Let $v \in \mathcal{B}_{\hat{T}_0M}(0, 7h/8)$ be fixed. Notice that $\pi^*(v) \in \mathcal{B}_{T_0M}(0, 7h/8)$. Hence, according to Lemma VI.14, there exists $x \in \mathcal{B}(0, h) \cap M$ such that $\pi^*(v) = \pi^*(x)$. We may write

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

where, since $\|\hat{T}_i\|_{op} \leq 1/h$, $\|R_k(v)\| \leq C_{k,d,\tau_{min},\mathbf{L}} \sqrt{f_{max}/f_{min}} h^k$ according to (VI.17). Then, according to (VI.18),

$$\begin{aligned}\pi^*(v) + \sum_{i=2}^{k-1} \hat{T}_i(\pi^*(v)^{\otimes i}) &= \pi^*(v) + \sum_{i=2}^{k-1} T_i^*(\pi^*(v)^{\otimes i}) + R'(\pi^*(v)) \\ &= \pi^*(x) + \sum_{i=2}^{k-1} T_i^*(\pi^*(x)^{\otimes i}) + R'(\pi^*(x)),\end{aligned}$$

where $\|R'(\pi^*(x))\| \leq C_{k,d,\tau_{min},L} \sqrt{f_{max}/f_{min}} h^{k+1}$. According to Lemma VI.14, we deduce that $\|\widehat{\Psi}(v) - x\| \leq C_{k,d,\tau_{min},\mathbf{L}} \sqrt{f_{max}/f_{min}} h^k$, hence

$$\sup_{u \in \hat{M}_{\texttt{POLY}}} d(u, M) \leq C_{k,d,\tau_{min},\mathbf{L}} \sqrt{\frac{f_{max}}{f_{min}}} h^k. \tag{VI.19}$$

Now we focus on $\sup_{x\in M} d(x, \hat{M}_{\texttt{POLY}})$. For this, we need a lemma ensuring that $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ covers M with high probability.

Lemma VI.20 (Lemma III.23)**.** *Let $h = \left(\frac{C_d' k}{f_{min}} \frac{\log n}{n}\right)^{1/d}$ with C_d' large enough. Then for n large enough so that $h \leq \tau_{min}/2$, with probability at least $1 - \left(\frac{1}{n}\right)^{k/d}$,*

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

Now we choose h satisfying the conditions of Proposition VI.15 and Lemma VI.20. Let x be in M and assume that $\|x - X_{j_0}\| \leq h$. According to (VI.18) and (VI.3), we deduce that $\|\widehat{\Psi}_{j_0}(\hat{\pi}_{j_0}(x)) - x\| \leq C_{k,d,\tau_{min},\mathbf{L}}\sqrt{f_{max}/f_{min}}h^k$. Hence, from Lemma VI.20,

$$\sup_{x\in M} d(x, \hat{M}_{\texttt{POLY}}) \leq C_{k,d,\tau_M,L}\sqrt{\frac{f_{max}}{f_{min}}}h^k \tag{VI.21}$$

with probability at least $1 - 2\left(\frac{1}{n}\right)^{k/d}$. Combining (VI.19) and (VI.21) gives Theorem VI.12.

VI.4.2 Minimax Lower Bounds

This section is devoted to describe the main ideas of the proofs of the minimax lower bounds, Theorems VI.9, VI.11 and VI.13. The methods we use rely on hypothesis comparison [Yu97]. We recall that for two distributions Q and Q' defined on the same space, the total variation distance $TV(Q, Q')$ and the L^1 test affinity $\|Q \wedge Q'\|_1$ are given by

$$TV(Q, Q') = \frac{1}{2}\int |dQ - dQ'|, \qquad \|Q \wedge Q'\|_1 = \int dQ \wedge dQ',$$

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

Le Cam's Lemma and Consequences

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

Theorem VI.22 (Le Cam's Lemma [Yu97])**.** *For all P, P' in the model $\mathcal{P}$,*

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

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

Moreover, $\|P^{\otimes n} \wedge P'^{\otimes n}\|_1 \geq \|P \wedge P'\|_1^n = (1 - TV(P, P'))^n$.

We derive Theorem VI.13, as well as Theorems VI.9 and VI.11 with fixed base point x, $\theta(P)$ being $supp(P) = M$, T_xM and $II_x^M \circ \pi_{T_xM}$ respectively. The hypotheses P, P' are built in Section VI.4.2. Such constructions are not substantially new in minimax geometric inference [GPPVW12a]. Therefore, we do not detail it further.

Conditional Assouad's Lemma

Now, consider the estimation of the differential quantities $T_{X_1}M$ and $II_{X_1}^M$ with random base point X_1. In both cases, the loss can be cast as

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

where $\hat{\theta} = \hat{\theta}(X, X')$, with $X = X_1$ driving the parameter of interest, and $X' = (X_2, \ldots, X_n) = X_{2:n}$. Since $\|.\|_{L^1(P)}$ obviously depends on P, the technique exposed in the previous section

does not apply anymore. However, a slight adaptation of Assouad's Lemma [Yu97] with an extra conditioning on $X = X_1$ carries out for our purpose. Let us now detail a general framework where the method applies.

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

$$\inf_{\hat{\theta}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \left[d(\theta_X(Q), \hat{\theta}(X, X')) \right], \tag{VI.23}$$

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

Given a set of probability distributions $\mathcal{Q}$ on $\mathcal{X} \times \mathcal{X}'$, write $\overline{Conv}(\mathcal{Q})$ for the set of mixture probability distributions with components in $\mathcal{Q}$. For all $\tau = (\tau_1, \dots, \tau_m) \in \{0,1\}^m$, τ^k denotes the m-tuple that differs from τ only at the kth position. We are now in position to state the conditional version of Assouad's Lemma that allows to lower bound the minimax risk (VI.23).

Lemma VI.24 (Conditional Assouad)**.** *Let $m \geq 1$ be an integer and let $\{\mathcal{Q}_\tau\}_{\tau \in \{0,1\}^m}$ be a family of 2^m submodels $\mathcal{Q}_\tau \subset \mathcal{Q}$. Let $\{U_k \times U'_k\}_{1 \leq k \leq m}$ be a family of pairwise disjoint subsets of $\mathcal{X} \times \mathcal{X}'$, and $\mathcal{D}_{\tau,k}$ be subsets of $\mathcal{D}$. Assume that for all $\tau \in \{0,1\}^m$ and $1 \leq k \leq m$,*

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

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

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

Then,

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

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

Notice that for a model of the form $\mathcal{Q} = \{\delta_{x_0} \otimes P, P \in \mathcal{P}\}$ with fixed $x_0 \in \mathcal{X}$, one recovers the classical Assouad's Lemma [Yu97] taking $U_k = \mathcal{X}$ and $U'_k = \mathcal{X}'$. Indeed, when $X = x$ a.s, the parameter of interest $\theta_X(Q) = \theta(Q)$ can be seen as non-random.

Construction of Hypotheses

In order to apply Le Cam's Lemma (Theorem VI.22) or the conditional Assouad's Lemma (Lemma VI.24), we describe in this section the construction of the hypotheses involved in the different contexts of estimation. For this, the strategy consists in building distributions that are stochastically close — i.e. with a large test affinity — for which the associated parameters of interest are as different as possible. Before continuing to the precise construction, let us make two remarks about the lower bounds with random point X_1.

First, the associated minimax risks (Theorems VI.9 and Theorem VI.11) involve the integration with respect to X_1. Hence, as for regression with L^p loss, multiple locations of bumps are required to yield the right rate. Second, building manifolds with different tangent spaces (resp curvature) would lead to locally singular distributions. Therefore it is natural to consider mixture distributions to get non-trivial bounds.

Let $M_0^{(0)}$ be a d-dimensional $\mathcal{C}^\infty$-submanifold of $\mathbb{R}^D$ with reach greater than 1 and such that it contains $\mathcal{B}_{\mathbb{R}^d\times\{0\}^{D-d}}(0,1/2)$. $M_0^{(0)}$ can be built for example by flattening smoothly a unit d-sphere. Since $M_0^{(0)}$ is $\mathcal{C}^\infty$, the uniform probability distribution $P_0^{(0)}$ on $M_0^{(0)}$ belongs to $\mathcal{P}^k_{1,\mathbf{L}^{(0)}}(1/V_0^{(0)}, 1/V_0^{(0)})$, for some $\mathbf{L}^{(0)}$ and $V_0^{(0)} = Vol(M_0^{(0)})$.

Let now $M_0 = (2\tau_{min})M_0^{(0)}$ be the submanifold obtained from $M_0^{(0)}$ by homothecy. By construction, from Proposition VI.5, we have $\tau_{M_0} \geq 2\tau_{min}$, $\mathcal{B}_{\mathbb{R}^d\times\{0\}^{D-d}}(0,\tau_{min}) \subset M_0$, and the uniform probability distribution P_0 on M_0 belongs to the model $\mathcal{P}^k_{2\tau_{min},\mathbf{L}}(f_{min}, f_{max})$ whenever $L_\perp \geq L_\perp^{(0)}/(2\tau_{min})$, ..., $L_k \geq L_k^{(0)}/(2\tau_{min})^{k-1}$, and provided that $f_{min} \leq \left((2\tau_{min})^d V_0^{(0)}\right)^{-1} \leq f_{max}$. Note that $L_\perp^{(0)}$, ..., $L_k^{(0)}$, $Vol(M_0^{(0)})$ depend only on d and k. For this reason, all the lower bounds will be valid for $\tau_{min}L_\perp, \ldots, \tau_{min}^{k-1}L_k, (\tau_{min}^d f_{min})^{-1}$ and $\tau_{min}^d f_{max}$ large enough to exceed the thresholds $L_\perp^{(0)}/2, \ldots, L_k^{(0)}/2^{k-1}$, $2^d V_0^{(0)}$ and $(2^d V_0^{(0)})^{-1}$ respectively.

For $0 < \delta \leq \tau_{min}/2$, let $x_1, \ldots, x_m \in M_0 \cap \mathcal{B}(0,\tau_{min})$ be such that for all $k \neq k'$, $\|x_k - x_{k'}\| \geq \delta$. A classical packing argument (see [Mas07] p. 71) shows that one can take up to $m = \lceil c_d/\delta^d \rceil$ for some $c_d > 0$. We let $e \in \mathbb{R}^D$ denote any unit vector orthogonal to $\mathbb{R}^d \times \{0\}^{D-d}$.

Let $\phi : \mathbb{R}^D \to [0,1]$ be a smooth scalar map such that $\phi|_{\mathcal{B}\left(0,\frac{1}{2}\right)} = 1$ and $\phi|_{\mathcal{B}(0,1)^c} = 0$. Let $\Lambda_+ > 0$ and $1 \geq A_+ > A_- > 0$ be real numbers to be chosen later. Let $\mathbf{\Lambda} = (\Lambda_1, \ldots, \Lambda_m)$ with entries $-\Lambda_+ \leq \Lambda_k \leq \Lambda_+$, and $\mathbf{A} = (A_1, \ldots, A_m)$ with entries $A_- \leq A_k \leq A_+$. For $z \in \mathbb{R}^D$, we write $z = (z_1, \ldots, z_D)$ for its coordinates in the canonical basis. For all $\tau = (\tau_1, \ldots, \tau_m) \in \{0,1\}^m$, define the bump map as

$$\Phi_\tau^{\mathbf{\Lambda},\mathbf{A},i}(x) = x + \sum_{k=1}^m \phi\left(\frac{x - x_k}{\delta}\right)\left\{\tau_k A_k (x - x_k)_1^i + (1-\tau_k)\Lambda_k\right\} e. \tag{VI.25}$$

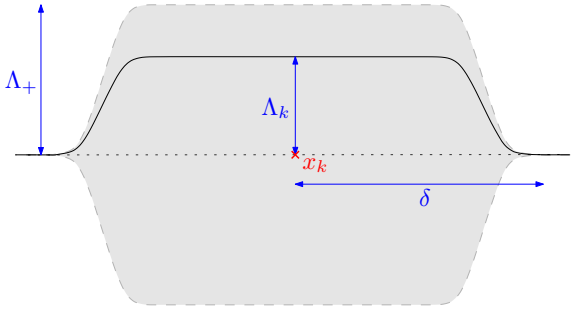
An analogous deformation map was considered in Section IV.4.1 to prove te interpolation Theorem IV.11. We let $P_\tau^{\mathbf{\Lambda},\mathbf{A},(i)}$ denote the pushforward distribution of P_0 by $\Phi_\tau^{\mathbf{\Lambda},\mathbf{A},(i)}$, and write $M_\tau^{\mathbf{\Lambda},\mathbf{A},(i)}$ for its support. Roughly speaking, $M_\tau^{\mathbf{\Lambda},\mathbf{A},i}$ consists of m bumps at the x_k's having different shapes (Figure VI.4). If $\tau_k = 0$, the bump at x_k is a symmetric plateau function and has height Λ_k. If $\tau_k = 1$, it fits the graph of the polynomial $A_k(x - x_k)_1^i$ locally. The following Lemma VI.26 gives differential bounds and geometric properties of $\Phi_\tau^{\mathbf{\Lambda},\mathbf{A},i}$. It follows straightforwardly from chain rule, similarly to Lemma IV.12.

Lemma VI.26. *There exists $c_{\phi,i} < 1$ such that if $A_+ \leq c_{\phi,i}\delta^{i-1}$ and $\Lambda_+ \leq c_{\phi,i}\delta$, then $\Phi_\tau^{\mathbf{\Lambda},\mathbf{A},i}$ is a global $\mathcal{C}^\infty$-diffeomorphism of $\mathbb{R}^D$ such that for all $1 \leq k \leq m$, $\Phi_\tau^{\mathbf{\Lambda},\mathbf{A},i}(\mathcal{B}(x_k,\delta)) = \mathcal{B}(x_k,\delta)$. Moreover,*

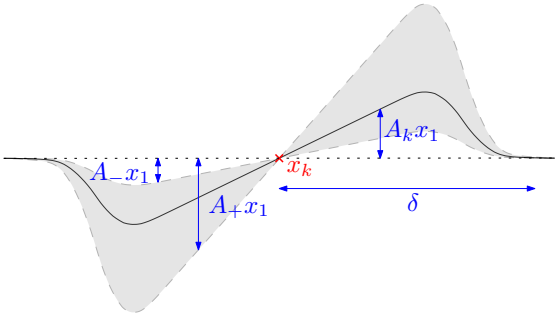
$$\left\|I_D - d\Phi_\tau^{\mathbf{\Lambda},\mathbf{A},i}\right\|_{op} \leq C_{\phi,i}\left\{\frac{A_+}{\delta^{1-i}}\right\} \vee \left\{\frac{\Lambda_+}{\delta}\right\},$$

and for $j \geq 2$,

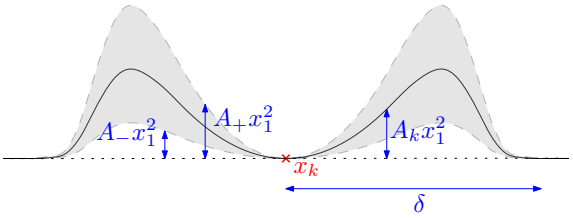
$$\left\|d^j\Phi_\tau^{\mathbf{\Lambda},\mathbf{A},i}\right\|_{op} \leq C_{\phi,i,j}\left\{\frac{A_+}{\delta^{j-i}}\right\} \vee \left\{\frac{\Lambda_+}{\delta^j}\right\}.$$



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



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



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

Figure VI.4 – The three shapes of the bump map $\Phi_\tau^{\mathbf{\Lambda},\mathbf{A},i}$ around a x_k.

Finally, we define the mixture distribution $\bar{Q}_{\tau,n}^{(i)}$ on $(\mathbb{R}^D)^n$ by

$$\bar{Q}_{\tau,n}^{(i)} = \int_{[-\Lambda_+,\Lambda_+]^m} \int_{[A_-,A_+]^m} \left(P_\tau^{\mathbf{\Lambda},\mathbf{A},(i)}\right)^{\otimes n} \frac{d\mathbf{A}}{(A_+ - A_-)^m} \frac{d\mathbf{\Lambda}}{(2\Lambda_+)^m}. \tag{VI.27}$$

Although the probability distribution $\bar{Q}_{\tau,n}^{(i)}$ depends on A_-, A_+ and Λ_+, we omit this dependency for the sake of compactness. Another way to define $\bar{Q}_{\tau,n}^{(i)}$ is the following: draw uniformly $\mathbf{\Lambda}$ in $[-\Lambda_+,\Lambda_+]^m$ and $\mathbf{A}$ in $[A_-,A_+]^m$, and given $(\mathbf{\Lambda},\mathbf{A})$, take $Z_i = \Phi_\tau^{\mathbf{\Lambda},\mathbf{A},i}(Y_i)$, where $Y_1,\dots,Y_n$ is an i.i.d. n-sample with common distribution P_0 on M_0. Then $(Z_1,\dots,Z_n)$ has distribution $\bar{Q}_{\tau,n}^{(i)}$.

We now state useful probabilistic and geometric properties of $\bar{Q}_{\tau,n}^{(i)}$, in view of using Theorem VI.24. For this, let us denote by $\mathcal{P}_\tau^{(i)}$ the set composed of all the distributions $P_\tau^{\mathbf{\Lambda},\mathbf{A},(i)}$ for $A_- \le A_1,\dots,A_m \le A_+$ and $-\Lambda_+ \le \Lambda_1,\dots,\Lambda_m \le \Lambda_+$. Again, we omit the dependency on A_-, A_+ and Λ_+.

Lemma VI.28. *Assume that the conditions of Lemma VI.26 hold, and let*

$$U_k = \mathcal{B}_{\mathbb{R}^d\times\{0\}^{D-d}}(x_k,\delta/2) + \mathcal{B}_{span(e)}(0,\tau_{min}/2),$$

where for $B, B' \subset \mathbb{R}^D$, $B + B'$ denotes their Minkovski sum, and

$$U_k' = \left(\mathbb{R}^D \setminus \left\{\mathcal{B}_{\mathbb{R}^d\times\{0\}^{D-d}}(x_k,\delta) + \mathcal{B}_{span(e)}(0,\tau_{min}/2)\right\}\right)^{n-1}.$$

Then the sets $U_k \times U_k'$ are pairwise disjoint, $\bar{Q}_{\tau,n}^{(i)} \in \overline{Conv}\left(\left(\mathcal{P}_\tau^{(i)}\right)^{\otimes n}\right)$, and if $(Z_1,\dots,Z_n) = (Z_1, Z_{2:n})$ has distribution $\bar{Q}_{\tau,n}^{(i)}$, Z_1 and $Z_{2:n}$ are independent conditionally on the event $\{(Z_1,Z_{2:n}) \in U_k \times U_k'\}$.

Moreover, if $(X_1,\dots,X_n)$ has distribution $\left(P_\tau^{\mathbf{\Lambda},\mathbf{A},(i)}\right)^{\otimes n}$ (with fixed $\mathbf{A}$ and $\mathbf{\Lambda}$), then on the event $\{X_1 \in U_k\}$, we have:

- *if* $\tau_k = 0$,

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

and $d_H(M_0, M_\tau^{\mathbf{\Lambda},\mathbf{A},(i)}) \geq |\Lambda_k|$.

- *if* $\tau_k = 1$,
  - *for* $i = 1$: $\angle\left(T_{X_1} M_\tau^{\mathbf{\Lambda},\mathbf{A},(1)}, \mathbb{R}^d \times \{0\}^{D-d}\right) \geq A_-/2$;
  - *for* $i = 2$: $\left\| II_{X_1}^{M_\tau^{\mathbf{\Lambda},\mathbf{A},(2)}} \circ \pi_{T_{X_1} M_\tau^{\mathbf{\Lambda},\mathbf{A},(2)}} \right\|_{op} \geq A_-/2$.

To apply Theorem VI.24 to the $\bar{Q}_{\tau,n}^{(i)}$'s with $\mathcal{X} = \mathbb{R}^D$, $\mathcal{X}' = \left(\mathbb{R}^D\right)^{n-1}$, it remains to bound the test affinities between their marginals on $\mathcal{X}$ and $\mathcal{X}'$. By construction (VI.27), these are respectively $\bar{Q}_{\tau,1}^{(i)}$ and $\bar{Q}_{\tau,n-1}^{(i)}$.

Lemma VI.29. *Assume that the conditions of Lemma VI.26 and Lemma VI.28 hold. If in addition,* $cA_+(\delta/4)^i \leq \Lambda_+ \leq CA_+(\delta/4)^i$ *for some absolute constants* $C \geq c > 3/4$, *and* $A_- = A_+/2$, *then,*

$$\int_{U_k} d\bar{Q}_{\tau,1}^{(i)} \wedge d\bar{Q}_{\tau^k,1}^{(i)} \geq \frac{c_{d,i}}{C} \left(\frac{\delta}{\tau_{min}}\right)^d,$$

and

$$\int_{U_k'} d\bar{Q}_{\tau,n-1}^{(i)} \wedge d\bar{Q}_{\tau^k,n-1}^{(i)} = \left(1 - c_d' \left(\frac{\delta}{\tau_{min}}\right)^d\right)^{n-1}.$$

Now, to derive Theorem VI.9, set $i = 1$, take $A_+ = 2A_- = \varepsilon\delta^{k-1}$, and $\Lambda_+ = \delta A_+/4$ for $\varepsilon = \varepsilon_{\phi,k,d,\tau_{min}}$ small enough so that $\mathcal{P}_\tau^{(1)} \subset \mathcal{P}^k_{\tau_{min},\mathbf{L}}(f_{min}, f_{max})$, according to Lemma VI.26 and Proposition VI.5. Hence, applying Lemma VI.24 together with Lemma VI.28 and Lemma VI.29, recalling that m can be taken of order c_d/δ^d, we get, for all estimators $\hat{T}$,

$$\begin{aligned}\sup_{P \in \mathcal{P}^k} \mathbb{E}_{P^{\otimes n}} \angle(T_{X_1} M, \hat{T}) &\geq c_{d,k} \varepsilon m \frac{A_-}{4} \left(\frac{\delta}{\tau_{min}}\right)^d \left(1 - c_d' \left(\frac{\delta}{\tau_{min}}\right)^d\right)^{n-1} \\ &\geq c'_{d,k,\tau_{min}} \frac{\delta^{k-1}}{\delta^d} \left(\frac{\delta}{\tau_{min}}\right)^d \left(1 - c_d' \left(\frac{\delta}{\tau_{min}}\right)^d\right)^{n-1}.\end{aligned}$$

Taking $(\delta/\tau_{min})^d = 1/(n-1)$ yields the result.

Similarly, to derive Theorem VI.11, set $i = 2$, take $A_+ = 2A_- = \varepsilon'\delta^{k-2}$, and $\Lambda_+ = \delta^2 A_+/4^2$ with $\varepsilon' = \varepsilon'_{\phi,k,d,\tau_{min}}$ small enough so that $\mathcal{P}_\tau^{(2)} \subset \mathcal{P}^k_{\tau_{min},\mathbf{L}}(f_{min}, f_{max})$. With $(\delta/\tau_{min})^d = 1/(n-1)$, the same derivation as above leads to the result.

Finally, for Theorem VI.13, simply take $m = 1$, $\tau = 0$ and $\Lambda_1 = \delta A_1/4 = \varepsilon\delta^k$ for $\varepsilon = \varepsilon_{\phi,k,d,\tau_{min}}$ as above. We may conclude using Theorem VI.22 with P_0 and $P_0^{\Lambda_1,A_1,(i)}$. Indeed, using $d_H(M_0, M_0^{\Lambda_1,A_1,(i)}) \geq |\Lambda_1| = \varepsilon\delta^k$ from Lemma VI.28, and noticing that the total variation distance between the two distributions is $P_0(\mathcal{B}(x_1,\delta)) = c_d(\delta/\tau_{min})^d$, since they differ only outside $\mathcal{B}(x_1,\delta)$, we get the result.

VI.5 Conclusion, Prospects

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

The implementation of (VI.7) needs to be investigated. The non-convexity of the criterion comes from that we minimize over the space of orthogonal projectors, which is non-convex. However, that space is pretty well understood, and it seems possible to implement gradient descents on it [UM14]. Another way to improve our procedure could be to fit orthogonal polynomials instead of monomials. Such a modification may also lead to improved dependency on the dimension d and the regularity k in the bounds for both tangent space and support estimation.

As a first attempt to a minimax study over models of higher order regularity $\mathcal{C}^k$ ($k \geq 3$) for submanifolds, we chose not to include noise. This is a limitation of the model $\mathcal{P}^k$, and one could argue that the methods described are not robust. However, with outliers in the model $\mathcal{C}^2$, we proposed in Chapter IV an iterative denoising procedure based on tangent space estimation. It exploits the fact that tangent space estimation allows to remove a part of outliers, and removing outliers enhances tangent space estimation. An interesting question would be to study how this method can apply with local polynomials.

Another open question is that of exact topology recovering with fast rates for $k \geq 3$. Indeed, $\hat{M}_{\texttt{POLY}}$ converges at rate $(\log n/n)^{k/d}$ but is unstructured. It would be nice to glue the patches of $\hat{M}_{\texttt{POLY}}$ together, for example using interpolation techniques, following the ideas of [FIK+15].

Appendix D

Proofs for Chapter VI

Content

D.1 Properties and Stability of the Models

D.1.1 Property of the Exponential Map in $\mathcal{C}^2_{\tau_{min}}$

Here we show the Lemma VI.1. From Proposition III.22 (ii) and (iii), we have that sectional curvatures of M satisfy $-2/\tau_{min}^2 \leq \kappa \leq 1/\tau_{min}^2$, and that the injectivity radius of M is at least $\pi\tau_{min} \geq \tau_{min}/4$. Therefore, $\exp_p : \mathcal{B}_{T_pM}(0, \tau_{min}/4) \to M$ is one-to-one.

Let us write $\mathbf{N}_p(v) = \exp_p(v) - p - v$. We clearly have $\mathbf{N}_p(0) = 0$ and $d_0\mathbf{N}_p = 0$. Let now $v \in \mathcal{B}_{T_pM}(0, \tau_{min}/4)$ be fixed. We have $d_v\mathbf{N}_p = d_v \exp_p - Id_{T_pM}$. For $0 \leq t \leq \|v\|$, we write $\gamma(t) = \exp_p(tv/\|v\|)$ for the arc-length parametrized geodesic from p to $\exp_p(v)$, and P_t for the parallel translation along γ. From Lemma 18 of [DVW15],

$$\left\| d_{t\frac{v}{\|v\|}} \exp_p - P_t \right\|_{op} \leq \frac{2}{\tau_{min}^2}\frac{t^2}{2} \leq \frac{t}{4\tau_{min}}.$$

We now derive an upper bound for $\left\|P_t - Id_{T_pM}\right\|_{op}$. For this, fix two unit vectors $u \in \mathbb{R}^D$ and $w \in T_pM$, and write $g(t) = \langle P_t(w) - w, u\rangle$. Letting $\bar{\nabla}$ denote the ambient derivative in $\mathbb{R}^D$, by definition of parallel translation,

$$\begin{aligned} |g'(t)| &= \left|\langle \bar{\nabla}_{\gamma'(t)} P_t(w) - w, u\rangle\right| \\ &= \left|\langle II^M_{\gamma(t)}(\gamma'(t), P_t(w)), u\rangle\right| \\ &\leq 1/\tau_{min}. \end{aligned}$$

Since $g(0) = 0$, we get $\left\|P_t - Id_{T_pM}\right\|_{op} \leq t/\tau_{min}$. Finally, the triangle inequality leads to

$$\begin{aligned}\left\|d_v\mathbf{N}_p\right\|_{op} &= \left\|d_v \exp - Id_{T_pM}\right\|_{op} \\ &\leq \left\|d_v \exp - P_{\|v\|}\right\|_{op} + \left\|P_{\|v\|} - Id_{T_pM}\right\|_{op} \\ &\leq \frac{5\left\|v\right\|}{4\tau_{min}}.\end{aligned}$$

We conclude with the property of the projection $\pi^* = \pi_{T_pM}$. Indeed, defining $R_2(y-p) = (y-p) - \pi^*(y-p)$, Lemma 4.7 in [Fed59] gives

$$\begin{aligned}\left\|R_2(y-p)\right\| &= d(y-p, T_pM) \\ &\leq \frac{\left\|y-p\right\|^2}{2\tau_{min}}.\end{aligned}$$

D.1.2 Geometric Properties of the $\mathcal{C}^k$ Models

We now move to the proof of Lemma VI.14.

Proof of Lemma VI.14. (i) Simply notice that from the reverse triangle inequality,

$$\left|\frac{\left\|\Psi_x(v_2) - \Psi_x(v_1)\right\|}{\left\|v_2 - v_1\right\|} - 1\right| \leq \frac{\left\|N_x(v_2) - N_x(v_1)\right\|}{\left\|v_2 - v_1\right\|} \leq L_\perp(\left\|v_1\right\| \vee \left\|v_2\right\|) \leq \frac{1}{4}.$$

(ii) The right-hand side inclusion follows straightforwardly from (i). Let us focus on the left-hand side inclusion. For this, consider the map defined by $G = \pi_{T_xM} \circ \Psi_x$ on the domain $\mathcal{B}_{T_xM}(0,h)$. For all $v \in \mathcal{B}_{T_xM}(0,h)$, we have

$$\left\|d_vG - Id_{T_xM}\right\|_{op} = \left\|\pi_{T_xM} \circ d_v\mathbf{N}_x\right\|_{op} \leq \left\|d_v\mathbf{N}_x\right\|_{op} \leq L_\perp \left\|v\right\| \leq \frac{1}{4} < 1.$$

Hence, G is a diffeomorphism onto its image and it satisfies $\left\|G(v)\right\| \geq 3\left\|v\right\|/4$. It follows that

$$\mathcal{B}_{T_xM}\left(0, \frac{3h}{4}\right) \subset G\left(\mathcal{B}_{T_xM}(0,h)\right) = \pi_{T_xM}\left(\Psi_x\left(\mathcal{B}_{T_xM}(0,h)\right)\right).$$

Now, according to Lemma VI.1, for all $y \in \mathcal{B}\left(x, \frac{3h}{5}\right) \cap M$,

$$\left\|\pi_{T_xM}(y-x)\right\| \leq \left\|y-x\right\| + \frac{\left\|y-x\right\|^2}{2\tau_{min}} \leq \left(1 + \frac{1}{4}\right)\left\|y-x\right\| \leq \frac{3h}{4},$$

from what we deduce $\pi_{T_xM}\left(\mathcal{B}\left(x, \frac{3h}{5}\right) \cap M\right) \subset \mathcal{B}_{T_xM}\left(0, \frac{3h}{4}\right)$. As a consequence,

$$\pi_{T_xM}\left(\mathcal{B}\left(x, \frac{3h}{5}\right) \cap M\right) \subset \pi_{T_xM}\left(\Psi_x\left(\mathcal{B}_{T_xM}(0,h)\right)\right),$$

which yields the announced inclusion since π_{T_xM} is one to one on $\mathcal{B}\left(x, \frac{5h}{4}\right) \cap M$ from Lemma 5 in [ACLZ17], and

$$\left(\mathcal{B}\left(x, \frac{3h}{5}\right) \cap M\right) \subset \Psi_x\left(\mathcal{B}_{T_xM}(0,h)\right) \subset \mathcal{B}\left(x, \frac{5h}{4}\right) \cap M.$$

(iii) Straightforward application of Lemma 5 in [ACLZ17].

(iv) Notice that Lemma VI.1 gives the existence of such an expansion for $k = 2$. Hence, we can assume $k \geq 3$. Taking $h = \frac{\tau_{min} \wedge L_\perp^{-1}}{4}$, we showed in the proof of (ii) that the map G is a diffeomorphism onto its image, with $\|d_v G - Id_{T_xM}\|_{op} \leq \frac{1}{4} < 1$. Additionally, the chain rule yields $\|d_v^i G\|_{op} \leq \|d_v^i \Psi_x\|_{op} \leq L_i$ for all $2 \leq i \leq k$. Therefore, from Lemma D.1, the differentials of G^{-1} up to order k are uniformly bounded. As a consequence, we get the announced expansion writing

$$y - x = \Psi_x \circ G^{-1}\left(\pi^*(y - x)\right),$$

and using the Taylor expansions of order k of Ψ_x and G^{-1}.

Let us now check that $T_2^* = II_x^M$. First, since by construction, T_2^* is the second order term of the Taylor expansion of $\Psi_x \circ G^{-1}$ at zero, a straightforward computation yields

$$\begin{aligned} T_2^* &= (I_D - \pi_{T_xM}) \circ d_0^2 \Psi_x \\ &= \pi_{T_xM^\perp} \circ d_0^2 \Psi_x. \end{aligned}$$

Let $v \in T_xM$ be fixed. Letting $\gamma(t) = \Psi_x(tv)$ for $|t|$ small enough, it is clear that $\gamma''(0) = d_0^2\Psi(v^{\otimes 2})$. Moreover, by definition of the second fundamental form [dC92, Proposition 2.1, p.127], since $\gamma(0) = x$ and $\gamma'(0) = v$, we have

$$II_x^M(v^{\otimes 2}) = \pi_{T_xM^\perp}(\gamma''(0)).$$

Hence

$$\begin{aligned} T_2^*(v^{\otimes 2}) &= \pi_{T_xM^\perp} \circ d_0^2\Psi_x(v^{\otimes 2}) \\ &= \pi_{T_xM^\perp}(\gamma''(0)) \\ &= II_x^M(v^{\otimes 2}), \end{aligned}$$

which concludes the proof.

(v) The first statement is a rephrasing of Proposition III.22 (i) and (ii).

□

In the proof of Lemma VI.14 (iv), we used a technical lemma of differential calculus that we now prove. It states quantitatively that if G is $\mathcal{C}^k$-close to the identity map, then it is a diffeomorphism onto its image and the differentials of its inverse G^{-1} are controlled.

Lemma D.1. *Let $k \geq 2$ and U be an open subset of $\mathbb{R}^d$. Let $G : U \to \mathbb{R}^d$ be $\mathcal{C}^k$. Assume that $\|I_d - dG\|_{op} \leq \varepsilon < 1$, and that for all $2 \leq i \leq k$, $\|d^i G\|_{op} \leq L_i$ for some $L_i > 0$. Then G is a $\mathcal{C}^k$-diffeomorphism onto its image, and for all $2 \leq i \leq k$,*

$$\left\|I_d - dG^{-1}\right\|_{op} \leq \frac{\varepsilon}{1-\varepsilon} \quad \text{and} \quad \left\|d^i G^{-1}\right\|_{op} \leq L'_{i,\varepsilon,L_2,\ldots,L_i} < \infty \text{ for } 2 \leq i \leq k.$$

Proof of Lemma D.1. For all $x \in U$, $\|d_x G - I_d\|_{op} < 1$, so G is one to one, and for all $y = G(x) \in G(U)$,

$$\begin{aligned} \left\|I_d - d_y G^{-1}\right\|_{op} &= \left\|I_d - (d_x G)^{-1}\right\|_{op} \\ &\leq \left\|(d_x G)^{-1}\right\|_{op} \|I_d - d_x G\|_{op} \\ &\leq \frac{\|I_d - d_x G\|_{op}}{1 - \|I_d - d_x G\|_{op}} \\ &\leq \frac{\varepsilon}{1-\varepsilon}. \end{aligned}$$

For $2 \leq i \leq k$ and $1 \leq j \leq i$, write $\Pi_i^{(j)}$ for the set of partitions of $\{1, \dots, i\}$ with j blocks. Differentiating i times the identity $G \circ G^{-1} = Id_{G(U)}$, Faa di Bruno's formula yields that, for all $y = G(x) \in G(U)$ and all unit vectors $h_1, \dots, h_i \in \mathbb{R}^D$,

$$0 = d_y \left(G \circ G^{-1} \right) . (h_\alpha)_{1 \leq \alpha \leq i} = \sum_{j=1}^{i} \sum_{\pi \in \Pi_i^{(j)}} d_x^j G . \left(\left(d_y^{|I|} G^{-1} . (h_\alpha)_{\alpha \in I} \right)_{I \in \pi} \right) .$$

Isolating the term for $j = 1$ entails

$$\begin{aligned}
&\left\| d_x \Phi . \left(d_y^i G^{-1} . (h_\alpha)_{1 \leq \alpha \leq i} \right) \right\|_{op} \\
&= \left\| - \sum_{j=2}^{i} \sum_{\pi \in \Pi_i^{(j)}} d_x^j G . \left(\left(d_y^{|I|} G^{-1} . (h_\alpha)_{\alpha \in I} \right)_{I \in \pi} \right) \right\|_{op} \\
&\leq \sum_{j=2}^{i} \sum_{\pi \in \Pi_i^{(j)}} \left\| d^j G \right\|_{op} \prod_{I \in \pi} \left\| d^{|I|} G^{-1} \right\|_{op} .
\end{aligned}$$

Using the first order Lipschitz bound on G^{-1}, we get

$$\left\| d^i G^{-1} \right\|_{op} \leq \frac{1 + \varepsilon}{1 - \varepsilon} \sum_{j=2}^{i} L_j \sum_{\pi \in \Pi_i^{(j)}} \prod_{I \in \pi} \left\| d^{|I|} G^{-1} \right\|_{op} .$$

The result follows by induction on i. □

D.1.3 Stability of the Models

This section is devoted to prove the stability of the model with respect to ambient diffeomorphisms (Proposition VI.5).

The second part is pretty straightforward since the dilation λM has reach $\tau_{\lambda M} = \lambda \tau_M$, and can be parametrized locally by $\tilde{\Psi}_{\lambda p}(v) = \lambda \Psi_p(v/\lambda) = \lambda p + v + \lambda \mathbf{N}_p(v/\lambda)$, yielding the differential bounds $\mathbf{L}_{(\lambda)}$. Bounds on the density follow from homogeneity of the d-dimensional Hausdorff measure.

For the first part, we split the proof into two intermediate results. Proposition D.2 deals with the stability of the geometric model, that is, the reach bound and the existence of a smooth parametrization when a submanifold is perturbed. Lemma D.3 deals with the condition on the density in the models $\mathcal{P}^k$. It gives a change of variable formula for pushforward of measure on submanifolds, ensuring a control on densities with respect to intrinsic volume measure.

Proposition D.2. *Let $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ be a global $\mathcal{C}^k$-diffeomorphism. If $\|d\Phi - I_D\|_{op}$, $\|d^2\Phi\|_{op}$, $\dots$, $\left\|d^k\Phi\right\|_{op}$ are small enough, then for all M in $\mathcal{C}^k_{\tau_{min},\mathbf{L}}$, the image $M' = \Phi(M)$ belongs to $\mathcal{C}^k_{\tau_{min}/2, 2L_\perp, 2L_3, \dots, 2L_k}$.*

Proof of Proposition D.2. To bound $\tau_{M'}$ from below, we use the stability of the reach with respect to $\mathcal{C}^2$ diffeomorphisms. Namely, from Theorem 4.19 in [Fed59] (see Lemma III.17),

$$\begin{aligned}
\tau_{M'} = \tau_{\Phi(M)} &\geq \frac{(1 - \|I_D - d\Phi\|_{op})^2}{\frac{1 + \|I_D - d\Phi\|_{op}}{\tau_M} + \|d^2\Phi\|_{op}} \\
&\geq \tau_{min} \frac{(1 - \|I_D - d\Phi\|_{op})^2}{1 + \|I_D - d\Phi\|_{op} + \tau_{min} \|d^2\Phi\|_{op}} \geq \frac{\tau_{min}}{2}
\end{aligned}$$

for $\|I_D - d\Phi\|_{op}$ and $\left\|d^2\Phi\right\|_{op}$ small enough. This shows the stability for $k = 2$, as well as that of the reach assumption for $k \geq 3$.

By now, take $k \geq 3$. We focus on the existence of a good parametrization of M' around a fixed point $p' = \Phi(p) \in M'$. For $v' \in T_{p'}M' = d_p\Phi\left(T_pM\right)$, let us define

$$\begin{aligned}\Psi'_{p'}(v') &= \Phi\left(\Psi_p\left(d_{p'}\Phi^{-1}.v'\right)\right)\\ &= p' + v' + \mathbf{N}'_{p'}(v'),\end{aligned}$$

where $\mathbf{N}'_{p'}(v') = \left\{\Phi\left(\Psi_p\left(d_{p'}\Phi^{-1}.v'\right)\right) - p' - v'\right\}$.

$$\begin{array}{ccc} M & \xrightarrow{\quad\Phi\quad} & M' \\ {\scriptstyle\Psi_p}\uparrow & & \uparrow{\scriptstyle\Psi'_{p'}} \\ T_pM & \xrightarrow[d_p\Phi]{} & T_{p'}M' \end{array}$$

The maps $\Psi'_{p'}(v')$ and $\mathbf{N}'_{p'}(v')$ are well defined whenever $\left\|d_{p'}\Phi^{-1}.v'\right\| \leq \frac{1}{8L_\perp}$, so in particular if $\|v'\| \leq \frac{1}{8(2L_\perp)} \leq \frac{1-\|I_D-d\Phi\|_{op}}{8L_\perp}$ and $\|I_D - d\Phi\|_{op} \leq \frac{1}{2}$. One easily checks that $\mathbf{N}'_{p'}(0) = 0$, $d_0\mathbf{N}'_{p'} = 0$ and writing $c(v') = p + d_{p'}\Phi^{-1}.v' + \mathbf{N}_{p'}\left(d_{p'}\Phi^{-1}.v'\right)$, for all unit vector $w' \in T_{p'}M'$,

$$\begin{aligned}\left\|d^2_{v'}\mathbf{N}'_{p'}(w'^{\otimes 2})\right\| &= \left\|d^2_{c(v')}\Phi\left(\left\{d_{d_{p'}\Phi^{-1}.v'}\Psi_p \circ d_{p'}\Phi^{-1}.w'\right\}^{\otimes 2}\right)\right.\\ &\qquad \left. + d_{c(v')}\Phi \circ d^2_{d_{p'}\Phi^{-1}.v'}\Psi_p\left(\left\{d_{p'}\Phi^{-1}.w'\right\}^{\otimes 2}\right)\right\|\\ &= \left\|d^2_{c(v')}\Phi\left(\left\{d_{d_{p'}\Phi^{-1}.v'}\Psi_p \circ d_{p'}\Phi^{-1}.w'\right\}^{\otimes 2}\right)\right.\\ &\qquad + \left(d_{c(v')}\Phi - Id\right) \circ d^2_{d_{p'}\Phi^{-1}.v'}\Psi_p\left(\left\{d_{p'}\Phi^{-1}.w'\right\}^{\otimes 2}\right)\\ &\qquad \left. + d^2_{d_{p'}\Phi^{-1}.v'}\Psi_p\left(\left\{d_{p'}\Phi^{-1}.w'\right\}^{\otimes 2}\right)\right\|\\ &\leq \left\|d^2\Phi\right\|_{op}\left(1 + L_\perp\left\|d_{p'}\Phi^{-1}.v'\right\|\right)^2\left\|d_{p'}\Phi^{-1}.w'\right\|^2\\ &\qquad + \|I_D - d\Phi\|_{op} L_\perp\left\|d_{p'}\Phi^{-1}.w'\right\|^2\\ &\qquad + L_\perp\left\|d_{p'}\Phi^{-1}.w'\right\|^2\\ &\leq \left\|d^2\Phi\right\|_{op}(1 + 1/8)^2\left\|d_{p'}\Phi^{-1}\right\|^2_{op}\\ &\qquad + \|I_D - d\Phi\|_{op} L_\perp\left\|d\Phi^{-1}\right\|^2_{op}\\ &\qquad + L_\perp\left\|d_{p'}\Phi^{-1}\right\|^2_{op}.\end{aligned}$$

Writing further $\left\|d\Phi^{-1}\right\|_{op} \leq (1 - \|I_D - d\Phi\|_{op})^{-1} \leq 1 + 2\|I_D - \Phi\|_{op}$ for $\|I_D - d\Phi\|_{op}$ small enough depending only on $L_\perp$, it is clear that the right-hand side of the latter inequality goes below $2L_\perp$ for $\|I_D - d\Phi\|_{op}$ and $\left\|d^2\Phi\right\|_{op}$ small enough. Hence, for $\|I_D - d\Phi\|_{op}$ and $\left\|d^2\Phi\right\|_{op}$ small enough depending only on $L_\perp$, $\|d^2_{v'}\mathbf{N}'_{p'}\|_{op} \leq 2L_\perp$ for all $\|v'\| \leq \frac{1}{8(2L_\perp)}$. From the chain rule, the same argument applies for the order $3 \leq i \leq k$ differential of $\mathbf{N}'_{p'}$. □

Lemma D.3 (Change of variable for the Hausdorff measure)**.** *Let P be a probability distribution on $M \subset \mathbb{R}^D$ with density f with respect to the d-dimensional Hausdorff*

measure $\mathcal{H}^d$. Let $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ be a global diffeomorphism such that $\|I_D - d\Phi\|_{\text{op}} < 1/3$. Let $P' = \Phi_* P$ be the pushforward of P by Φ. Then P' has a density g with respect to $\mathcal{H}^d$. This density can be chosen to be, for all $z \in \Phi(M)$,

$$g(z) = \frac{f\left(\Phi^{-1}(z)\right)}{\sqrt{\det\left(\pi_{T_{\Phi^{-1}(z)}M} \circ d_{\Phi^{-1}(z)}\Phi^T \circ \restriction d_{\Phi^{-1}(z)}\Phi T_{\Phi^{-1}(z)}M\right)}}.$$

In particular, if $f_{min} \leq f \leq f_{max}$ on M, then for all $z \in \Phi(M)$,

$$\left(1 - 3d/2\,\|I_D - d\Phi\|_{\text{op}}\right) f_{min} \leq g(z) \leq f_{max}\left(1 + 3(2^{d/2} - 1)\,\|I_D - d\Phi\|_{\text{op}}\right).$$

Proof of Lemma D.3. Let $p \in M$ be fixed and $A \subset \mathcal{B}(p, r) \cap M$ for r small enough. For a differentiable map $h : \mathbb{R}^d \to \mathbb{R}^D$ and for all $x \in \mathbb{R}^d$, we let $J_h(x)$ denote the d-dimensional Jacobian $J_h(x) = \sqrt{\det\left(d_x h^T\, d_x h\right)}$. The area formula ([Fed69, Theorem 3.2.5]) states that if h is one-to-one,

$$\int_A u\left(h(x)\right) J_h(x) \lambda^d(dx) = \int_{h(A)} u(y) \mathcal{H}^d(dy),$$

whenever $u : \mathbb{R}^D \to \mathbb{R}$ is Borel, where λ^d is the Lebesgue measure on $\mathbb{R}^d$. By definition of the pushforward, and since $dP = f d\mathcal{H}^d$,

$$\int_{\Phi(A)} dP'(z) = \int_A f(y) \mathcal{H}^d(dy).$$

Writing $\Psi_p = \exp_p : T_p M \to \mathbb{R}^D$ for the exponential map of M at p, we have

$$\int_A f(y) \mathcal{H}^d(dy) = \int_{{\Psi_p}^{-1}(A)} f(\Psi_p(x)) J_{\Psi_p}(x) \lambda^d(dx).$$

Rewriting the right hand term, we apply the area formula again with $h = \Phi \circ \Psi_p$,

$$\begin{aligned}
&\int_{{\Psi_p}^{-1}(A)} f(\Psi_p(x)) J_{\Psi_p}(x) \lambda^d(dy) \\
&\qquad = \int_{{\Psi_p}^{-1}(A)} f\left(\Phi^{-1}\left(h(x)\right)\right) \frac{J_{\Psi_p}\left(h^{-1}\left(h(x)\right)\right)}{J_{\Phi\circ\Psi_p}\left(h^{-1}\left(h(x)\right)\right)} J_{\Phi\circ\Psi_p}(x) \lambda^d(dx) \\
&\qquad = \int_{\Phi(A)} f\left(\Phi^{-1}(z)\right) \frac{J_{\Psi_p}\left(h^{-1}(z)\right)}{J_{\Phi\circ\Psi_p}\left(h^{-1}(z)\right)} \mathcal{H}^d(dz).
\end{aligned}$$

Since this is true for all $A \subset \mathcal{B}(p, r) \cap M$, P' has a density g with respect to $\mathcal{H}^d$, with

$$g(z) = f\left(\Phi^{-1}(z)\right) \frac{J_{\Psi_{\Phi^{-1}(z)}}\left(\Psi^{-1}_{\Phi^{-1}(z)} \circ \Phi^{-1}(z)\right)}{J_{\Phi\circ\Psi_{\Phi^{-1}(z)}}\left(\Psi^{-1}_{\Phi^{-1}(z)} \circ \Phi^{-1}(z)\right)}.$$

Writing $p = \Phi^{-1}(z)$, it is clear that $\Psi^{-1}_{\Phi^{-1}(z)} \circ \Phi^{-1}(z) = \Psi_p^{-1}(p) = 0 \in T_p M$. Since $d_0 \exp_p : T_p M \to \mathbb{R}^D$ is the inclusion map, we get the first statement.

We now let B and π_T denote $d_p\Phi$ and $\pi_{T_p M}$ respectively. For any unit vector $v \in T_p M$,

$$\begin{aligned}
\left|\left\|\pi_T B^T B v\right\| - \|v\|\right| &\leq \left\|\pi_T\left(B^T B - I_D\right) v\right\| \\
&\leq \left\|B^T B - I_D\right\|_{\text{op}} \\
&\leq \left(2 + \|I_D - B\|_{\text{op}}\right) \|I_D - B\|_{\text{op}} \\
&\leq 3\,\|I_D - B\|_{\text{op}}.
\end{aligned}$$

Therefore, $1 - 3\left\|I_D - B\right\|_{\mathrm{op}} \leq \left\|\pi_T B^T \upharpoonright BT_pM\right\|_{\mathrm{op}} \leq 1 + 3\left\|I_D - B\right\|_{\mathrm{op}}$. Hence,

$$\sqrt{\det\left(\pi_T B^T \upharpoonright BT_pM\right)} \leq \left(1 + 3\left\|I_D - B\right\|_{\mathrm{op}}\right)^{d/2} \leq \frac{1}{1 - \frac{3d}{2}\left\|I_D - B\right\|_{\mathrm{op}}},$$

and

$$\sqrt{\det\left(\pi_T B^T \upharpoonright BT_pM\right)} \geq \left(1 - 3\left\|I_D - B\right\|_{\mathrm{op}}\right)^{d/2} \geq \frac{1}{1 + 3(2^{d/2} - 1)\left\|I_D - B\right\|_{\mathrm{op}}},$$

which yields the result. □

D.2 Some Probabilistic Tools

D.2.1 Volume and Covering Rate

The first lemma of this section gives some details about the covering rate of a manifold with bounded reach.

Lemma D.4 (Lemma III.23)**.** *Let $P \in \mathcal{P}^k$ have support $M \subset \mathbb{R}^D$. Then for all $r \leq \tau_{min}/4$ and x in M,*

$$c_d f_{min} r^d \leq p_x(r) \leq C_d f_{max} r^d,$$

for some $c_d, C_d > 0$, with $p_x(r) = P(\mathcal{B}(x,r))$. Moreover, letting $h = \left(\frac{C'_d k}{f_{min}} \frac{\log n}{n}\right)^{1/d}$ with C'_d large enough, the following holds. For n large enough so that $h \leq \tau_{min}/2$, with probability at least $1 - \left(\frac{1}{n}\right)^{k/d}$,

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

D.2.2 Concentration Bounds for Local Polynomials

This section is devoted to the proof of Proposition VI.15. A first step is to ensure that empirical expectations order k polynomials are close to their deterministic counterparts.

Proposition D.5. *For any $x \in M$, we have*

$$\mathbb{P}\left[\sup_{u_1,\ldots,u_k,\varepsilon \in \{0,1\}^k} \left|(P - P_{n-1}) \prod_{j=1}^{p} \left(\frac{\langle u_j, y\rangle}{h}\right)^{\varepsilon_j} \mathbb{1}_{\mathcal{B}(x,h)}(y)\right| \right.$$
$$\left. \geq p_x(h)\left(\frac{4k\sqrt{2\pi}}{\sqrt{(n-1)p_x(h)}} + \sqrt{\frac{2t}{(n-1)p_x(h)}} + \frac{2}{3(n-1)p_x(h)}\right)\right] \leq e^{-t},$$

where P_{n-1} denotes the empirical distribution of $n-1$ i.i.d. random variables X_i drawn from P.

Proof of Proposition D.5. Without loss of generality we choose $x = 0$ and shorten notation to $\mathcal{B}(h)$ and $p(h)$. Let Z denote the empirical process on the left-hand side of Proposition D.5. Denote also by $f_{u,\varepsilon}$ the map $\prod_{j=1}^{k}\left(\frac{\langle u_j, y\rangle}{h}\right)^{\varepsilon_j} \mathbb{1}_{\mathcal{B}(h)}(y)$, and let $\mathcal{F}$ denote the set of such maps, for u_j in $\mathcal{B}(1)$ and ε in $\{0,1\}^k$.

Since $\|f_{u,\varepsilon}\|_\infty \le 1$ and $Pf_{u,\varepsilon}^2 \le p(h)$, the Talagrand-Bousquet inequality ([Bou02, Theorem 2.3]) yields

$$Z \le 4\mathbb{E}Z + \sqrt{\frac{2p(h)t}{n-1}} + \frac{2t}{3(n-1)},$$

with probability larger than $1 - e^{-t}$. It remains to bound $\mathbb{E}Z$ from above.

Lemma D.6. *We may write*

$$\mathbb{E}Z \le \frac{\sqrt{2\pi p(h)}}{\sqrt{n-1}}k.$$

Proof of Lemma D.6. Let σ_i and g_i denote some independent Rademacher and Gaussian variables. For convenience, we denote by $\mathbb{E}_A$ the expectation with respect to the random variable A. Using symmetrization inequalities we may write

$$\begin{aligned}
\mathbb{E}Z &= \mathbb{E}_X \sup_{u,\varepsilon} \left| (P - P_{n-1}) \prod_{j=1}^k \left(\frac{\langle u_j, y\rangle}{h} \right)^{\varepsilon_j} \mathbb{1}_{\mathcal{B}(h)}(y) \right| \\
&\le \frac{2}{n-1} \mathbb{E}_X \mathbb{E}_\sigma \sup_{u,\varepsilon} \sum_{i=1}^{n-1} \sigma_i \prod_{j=1}^k \left(\frac{\langle u_j, X_i\rangle}{h} \right)^{\varepsilon_j} \mathbb{1}_{\mathcal{B}(h)}(X_i) \\
&\le \frac{\sqrt{2\pi}}{n-1} \mathbb{E}_X \mathbb{E}_g \sup_{u,\varepsilon} \sum_{i=1}^{n-1} g_i \prod_{j=1}^k \left(\frac{\langle u_j, X_i\rangle}{h} \right)^{\varepsilon_j} \mathbb{1}_{\mathcal{B}(h)}(X_i).
\end{aligned}$$

Now let Y_g denote the Gaussian process $\sum_{i=1}^{n-1} g_i \prod_{j=1}^k \left(\frac{\langle u_j, X_i\rangle}{h} \right)^{\varepsilon_j} \mathbb{1}_{\mathcal{B}(h)}(X_i)$. Since, for any x in $\mathcal{B}(h)$, u,v in $\mathcal{B}(1)^k$, and ε, ε' in $\{0,1\}^k$, we have

$$\begin{aligned}
&\left| \prod_{j=1}^k \left(\frac{\langle x, u_j\rangle}{h} \right)^{\varepsilon_j} - \prod_{j=1}^k \left(\frac{\langle x, v_j\rangle}{h} \right)^{\varepsilon'_j} \right| \\
&\qquad \le \left| \sum_{r=1}^k \left(\prod_{j=1}^{k+1-r} \left(\frac{\langle x, u_j\rangle}{h} \right)^{\varepsilon_j} \prod_{j=k+2-r}^k \left(\frac{\langle x, v_j\rangle}{h} \right)^{\varepsilon'_j} \right.\right. \\
&\qquad\qquad \left.\left. - \prod_{j=1}^{k-r} \left(\frac{\langle x, u_j\rangle}{h} \right)^{\varepsilon_j} \prod_{j=k+1-r}^k \left(\frac{\langle x, v_j\rangle}{h} \right)^{\varepsilon'_j} \right) \right| \\
&\qquad \le \sum_{r=1}^k \left| \prod_{j=1}^{k-r} \left(\frac{\langle x, u_j\rangle}{h} \right)^{\varepsilon_j} \prod_{j=k+2-r}^k \left(\frac{\langle x, v_j\rangle}{h} \right)^{\varepsilon'_j} \left[\left(\frac{\langle u_{k+1-r}, x\rangle}{h} \right)^{\varepsilon_{k+1-r}} \right.\right. \\
&\qquad\qquad \left.\left. - \left(\frac{\langle v_{k+1-r}, x\rangle}{h} \right)^{\varepsilon'_{k+1-r}} \right] \right| \\
&\qquad \le \sum_{r=1}^k \left| \frac{\langle \varepsilon_r u_r - \varepsilon'_r v_r, x\rangle}{h} \right|.
\end{aligned}$$

We deduce that

$$\begin{aligned}
\mathbb{E}_g (Y_{u,\varepsilon} - Y_{v,\varepsilon'})^2 &\le k \sum_{i=1}^{n-1} \sum_{r=1}^k \left(\frac{\langle \varepsilon_r u_r, X_i\rangle}{h} - \frac{\langle \varepsilon'_r v_r, X_i\rangle}{h} \right)^2 \mathbb{1}_{\mathcal{B}(h)}(X_i) \\
&\le \mathbb{E}_g (\Theta_{u,\varepsilon} - \Theta_{v,\varepsilon'})^2,
\end{aligned}$$

where $\Theta_{u,\varepsilon} = \sqrt{k}\sum_{i=1}^{n-1}\sum_{r=1}^{k} g_{i,r}\frac{\langle \varepsilon_r u_r, X_i\rangle}{h}\mathbb{1}_{\mathcal{B}(h)}(X_i)$. According to Slepian's Lemma [BLM13, Theorem 13.3], it follows that

$$\begin{aligned}\mathbb{E}_g \sup_{u,\varepsilon} Y_g &\leq \mathbb{E}_g \sup_{u,\varepsilon} \Theta_g \\ &\leq \sqrt{k}\mathbb{E}_g \sup_{u,\varepsilon} \sum_{r=1}^{k} \frac{\left\langle \varepsilon_r u_r, \sum_{i=1}^{n-1} g_{i,r}\mathbb{1}_{\mathcal{B}(h)}(X_i)X_i\right\rangle}{h} \\ &\leq \sqrt{k}\mathbb{E}_g \sup_{u,\varepsilon} \sqrt{k\sum_{r=1}^{k} \frac{\left\langle \varepsilon_r u_r, \sum_{i=1}^{n-1} g_{i,r}\mathbb{1}_{\mathcal{B}(h)}(X_i)X_i\right\rangle^2}{h^2}}.\end{aligned}$$

We deduce that

$$\begin{aligned}\mathbb{E}_g \sup_{u,\varepsilon} Y_g &\leq \mathbb{E}_g \sup_{u,\varepsilon} \Theta_g \\ &\leq k\sqrt{\mathbb{E}_g \sup_{\|u\|=1,\varepsilon\in\{0,1\}} \frac{\left\langle \varepsilon u, \sum_{i=1}^{n-1} g_i\mathbb{1}_{\mathcal{B}(h)}(X_i)X_i\right\rangle^2}{h^2}} \\ &\leq k\sqrt{\mathbb{E}_g \left\|\sum_{i=1}^{n-1} \frac{g_i X_i}{h}\mathbb{1}_{\mathcal{B}(h)}(X_i)\right\|^2} \\ &\leq k\sqrt{N(h)}.\end{aligned}$$

Then we can deduce that $\mathbb{E}_X\mathbb{E}_g \sup_{u,\varepsilon} Y_g \leq k\sqrt{p(h)}$, hence the result. □

Combining Lemma D.6 with Talagrand-Bousquet's inequality gives the result of Proposition D.5. □

We are now in position to prove Proposition VI.15.

Proof of Proposition VI.15. If $h \leq \tau_{min}/8$, then, according to Lemma D.4, $p(h) \geq c_d f_{min} h^d$, hence, if $h = \left(K\frac{\log(n)}{n-1}\right)^{\frac{1}{d}}$, $(n-1)p(h) \geq K c_d f_{min}\log(n)$. Choosing $t = (k/d+1)\log(n)$ in Proposition D.5 and $K = K'/f_{min}$, with $K' > 1$ leads to

$$\mathbb{P}\left[\sup_{u_1,\dots,u_k,\varepsilon\in\{0,1\}^k} \left|(P-P_{n-1})\prod_{j=1}^{k}\left(\frac{\langle u_j, y\rangle}{h}\right)^{\varepsilon_j}\mathbb{1}_{\mathcal{B}(x,h)}(y)\right| \geq \frac{c_{d,k}f_{max}}{\sqrt{K'}}h^d\right] \leq \left(\frac{1}{n}\right)^{\frac{k}{d}+1}.$$

On the complement of the probability event mentioned just above, for a polynomial $Q = \sum_{\alpha\in[0,k]^d||\alpha|\leq k} a_\alpha x_{1:d}^\alpha$, we have

$$\begin{aligned}(P_{n-1}-P)Q^2(x_{1:d})\mathbb{1}_{\mathcal{B}(h)}(x) &\geq -\sum_{\alpha,\beta}\frac{c_{d,k}f_{max}}{\sqrt{K'}}|a_\alpha a_\beta|h^{d+|\alpha|+|\beta|} \\ &\geq -\frac{c_{d,k}f_{max}}{\sqrt{K'}}h^d\|Q_h\|_2^2.\end{aligned}$$

On the other hand, we may write, for all $r > 0$,

$$\int_{\mathcal{B}(0,r)} Q^2(x_{1:d})dx_1\dots dx_d \geq C_{d,k}r^d\|Q_r\|_2^2,$$

for some constant $C_{d,k}$. It follows that

$$PQ^2(x_{1:d})\mathbb{1}_{\mathcal{B}(h)}(x) \geq PQ^2(x_{1:d})\mathbb{1}_{B(7h/8)}(x_{1:d}) \geq c_{k,d}h^d f_{min}\|Q_h\|_2^2,$$

according to Lemma VI.14. Then we may choose $K' = \kappa_{k,d}(f_{max}/f_{min})^2$, with $\kappa_{k,d}$ large enough so that

$$P_{n-1}Q^2(x_{1:d})\mathbb{1}_{\mathcal{B}(h)}(x) \geq c_{k,d}f_{min}h^d\|Q_h\|_2^2.$$

□

D.3 Minimax Lower Bounds

D.3.1 Proof of the Conditional Assouad's Lemma

This section is dedicated to the proof of Lemma VI.24. The proof follows that of Lemma 2 in [Yu97]. Let $\hat{\theta} = \hat{\theta}(X, X')$ be fixed. For any family of 2^m distributions $\{Q_\tau\}_\tau \in \{\mathcal{Q}_\tau\}_\tau$, since the $U_k \times U_k'$'s are pairwise disjoint,

$$\begin{aligned}
&\sup_{Q\in\mathcal{Q}} \mathbb{E}_Q\left[d(\theta_X(Q), \hat{\theta}(X, X'))\right] \\
&\quad\geq \max_\tau \mathbb{E}_{Q_\tau} d(\hat{\theta}, \theta_X(Q_\tau)) \\
&\quad\geq \max_\tau \mathbb{E}_{Q_\tau} \sum_{k=1}^m d(\hat{\theta}, \theta_X(Q_\tau))\mathbb{1}_{U_k\times U_k'}(X, X') \\
&\quad\geq 2^{-m}\sum_\tau \sum_{k=1}^m \mathbb{E}_{Q_\tau} d(\hat{\theta}, \theta_X(Q_\tau))\mathbb{1}_{U_k\times U_k'}(X, X') \\
&\quad\geq 2^{-m}\sum_\tau \sum_{k=1}^m \mathbb{E}_{Q_\tau} d(\hat{\theta}, \mathcal{D}_{\tau,k})\mathbb{1}_{U_k\times U_k'}(X, X') \\
&\quad= \sum_{k=1}^m 2^{-(m+1)} \sum_\tau \Big(\mathbb{E}_{Q_\tau} d(\hat{\theta}, \mathcal{D}_{\tau,k})\mathbb{1}_{U_k\times U_k'}(X, X') + \mathbb{E}_{Q_{\tau^k}} d(\hat{\theta}, \mathcal{D}_{\tau^k,k})\mathbb{1}_{U_k\times U_k'}(X, X')\Big).
\end{aligned}$$

Since the previous inequality holds for all $Q_\tau \in \mathcal{Q}_\tau$, it extends to $\overline{Q}_\tau \in \overline{Conv}(\mathcal{Q}_\tau)$ by linearity. Let us now lower bound each of the terms of the sum for fixed $\tau \in \{0,1\}^m$ and $1 \leq k \leq m$. By assumption, if (X, X') has distribution $\overline{Q}_\tau$, then conditionally on $\{(X, X') \in U_k \times U_k'\}$, X and X' are independent. Therefore,

$$\begin{aligned}
&\mathbb{E}_{\overline{Q}_\tau} d(\hat{\theta}, \mathcal{D}_{\tau,k})\mathbb{1}_{U_k\times U_k'}(X, X') + \mathbb{E}_{\overline{Q}_{\tau^k}} d(\hat{\theta}, \mathcal{D}_{\tau^k,k})\mathbb{1}_{U_k\times U_k'}(X, X') \\
&\quad\geq \mathbb{E}_{\overline{Q}_\tau} d(\hat{\theta}, \mathcal{D}_{\tau,k})\mathbb{1}_{U_k}(X)\mathbb{1}_{U_k'}(X') + \mathbb{E}_{\overline{Q}_{\tau^k}} d(\hat{\theta}, \mathcal{D}_{\tau^k,k})\mathbb{1}_{U_k}(X)\mathbb{1}_{U_k'}(X') \\
&\quad= \mathbb{E}_{\bar{\nu}_\tau}\left[\mathbb{E}_{\bar{\mu}_\tau}\left(d(\hat{\theta}, \mathcal{D}_{\tau,k})\mathbb{1}_{U_k}(X)\right)\mathbb{1}_{U_k'}(X')\right] \\
&\qquad\quad + \mathbb{E}_{\bar{\nu}_{\tau^k}}\left[\mathbb{E}_{\bar{\mu}_{\tau^k}}\left(d(\hat{\theta}, \mathcal{D}_{\tau^k,k})\mathbb{1}_{U_k}(X)\right)\mathbb{1}_{U_k'}(X')\right] \\
&\quad= \int_{U_k}\int_{U_k'} d(\hat{\theta}, \mathcal{D}_{\tau,k})d\bar{\mu}_\tau(x)d\bar{\nu}_\tau(x') + \int_{U_k}\int_{U_k'} d(\hat{\theta}, \mathcal{D}_{\tau^k,k})d\bar{\mu}_{\tau^k}(x)d\bar{\nu}_{\tau^k}(x') \\
&\quad\geq \int_{U_k}\int_{U_k'} \left(d(\hat{\theta}, \mathcal{D}_{\tau,k}) + d(\hat{\theta}, \mathcal{D}_{\tau^k,k})\right) d\bar{\mu}_\tau \wedge d\bar{\mu}_{\tau^k}(x)d\bar{\nu}_\tau \wedge d\bar{\nu}_{\tau^k}(x') \\
&\quad\geq \Delta\left(\int_{U_k} d\bar{\mu}_\tau \wedge d\bar{\mu}_{\tau^k}\right)\left(\int_{U_k'} d\bar{\nu}_\tau \wedge d\bar{\nu}_{\tau^k}\right) \\
&\quad\geq \Delta(1-\alpha),
\end{aligned}$$

where we used that $d(\hat{\theta}, \mathcal{D}_{\tau,k}) + d(\hat{\theta}, \mathcal{D}_{\tau^k,k}) \geq \Delta$. The result follows by summing the bound above $|\{1,\dots,m\} \times \{0,1\}^m| = m2^m$ times.

D.3.2 Construction of Generic Hypotheses

In this section we prove Lemma VI.28 and Lemma VI.29.

Proof of Lemma VI.28. It is clear from the definition (VI.27) that $\bar{Q}^{(i)}_{\tau,n} \in \overline{Conv}((\mathcal{P}^{(i)}_\tau)^{\otimes n})$. By construction of the $\Phi^{\mathbf{\Lambda},\mathbf{A},i}_\tau$'s, these maps leave the sets

$$\mathcal{B}_{\mathbb{R}^d\times\{0\}^{D-d}}(x_k,\delta) + \mathcal{B}_{span(e)}(0,\tau_{min}/2)$$

unchanged for all $\mathbf{\Lambda},\mathbf{L}$. Therefore, on the event $\{(Z_1,Z_{2:n}) \in U_k \times U'_k\}$, one can write Z_1 only as a function of X_1,Λ_k,A_k, and $Z_{2:n}$ as a function of the rest of the X_j's,Λ_k's and A_k's. Therefore, Z_1 and $Z_{2:n}$ are independent.

We now focus on the geometric statements. For this, we fix a deterministic point $z = \Phi^{\mathbf{\Lambda},\mathbf{A},(i)}_\tau(x_0) \in U_k \cap M^{\mathbf{\Lambda},\mathbf{A},(i)}_\tau$. By construction, one necessarily has $x_0 \in M_0 \cap \mathcal{B}(x_k,\delta/2)$.

- If $\tau_k = 0$, locally around x_0, $\Phi^{\mathbf{\Lambda},\mathbf{A},(1)}_\tau$ is the translation of vector $\Lambda_k e$. Therefore, since M_0 satisfies $T_{x_0}M_0 = \mathbb{R}^d \times \{0\}^{D-d}$ and $II^{M_0}_{x_0} = 0$, we have

$$T_z M^{\mathbf{\Lambda},\mathbf{A},(i)}_\tau = \mathbb{R}^d \times \{0\}^{D-d} \quad \text{and} \quad \left\| II^{M^{\mathbf{\Lambda},\mathbf{A},(i)}_\tau}_z \circ \pi_{T_z M^{\mathbf{\Lambda},\mathbf{A},(i)}_\tau} \right\|_{op} = 0.$$

  Furthermore, by construction, $z_k = x_k + \Lambda_k e$ belongs to $M^{\mathbf{\Lambda},\mathbf{A},(i)}_\tau$. Since e is orthogonal to M_0, $d(z_0,M_0) \geq |\Lambda_k|$. Thus

$$d_H\big(M_0, M^{\mathbf{\Lambda},\mathbf{A},(i)}_\tau\big) \geq |\Lambda_k|.$$

- if $\tau_k = 1$,

  - for $i=1$: locally around x_0, $\Phi^{\mathbf{\Lambda},\mathbf{A},(1)}_\tau$ can be written as $x \mapsto x + A_k(x-x_k)_1 e$. Hence, $T_z M^{\mathbf{\Lambda},\mathbf{A},(i)}_\tau$ contains the direction $(1,A_k)$ in the plane $span(e_1,e)$ spanned by the first vector of the canonical basis and e. As a consequence, since e is orthogonal to $\mathbb{R}^d \times \{0\}^{D-d}$,

$$\angle\left(T_z M^{\mathbf{\Lambda},\mathbf{A},(1)}_\tau, \mathbb{R}^d \times \{0\}^{D-d}\right) \geq \left(1 + 1/A_k^2\right)^{-1/2} \geq A_k/2 \geq A_-/2.$$

  - for $i=2$: locally around x_0, $\Phi^{\mathbf{\Lambda},\mathbf{A},(2)}_\tau$ can be written as $x \mapsto x + A_k(x-x_k)_1^2 e$. Hence, $M^{\mathbf{\Lambda},\mathbf{A},(2)}_\tau$ contains an arc of parabola of equation $y = A_k(x-x_k)_1^2$ in the plane $span(e_1,e)$. As a consequence,

$$\left\| II^{M^{\mathbf{\Lambda},\mathbf{A},(2)}_\tau}_z \circ \pi_{T_z M^{\mathbf{\Lambda},\mathbf{A},(2)}_\tau} \right\|_{op} \geq A_k/2 \geq A_-/2.$$

□

Proof of Lemma VI.29. First note that all the distributions involved have support in $\mathbb{R}^d \times span(e) \times \{0\}^{D-(d+1)}$. Therefore, we use the canonical coordinate system of $\mathbb{R}^d \times span(e)$, centered at x_k, and we denote the components by $(x_1,x_2,\ldots,x_d,y) = (x_1,x_{2:d},y)$. Without loss of generality, assume that $\tau_k = 0$ (if not, flip τ and τ^k). Recall that ϕ has been chosen to be constant and equal to 1 on the ball $\mathcal{B}(0,1/2)$.

By definition (VI.27), on the event $\{Z \in U_k\}$, a random variable Z having distribution $\bar{Q}^{(i)}_{\tau,1}$ can be represented as $Z = X + \phi\left(\frac{X-x_k}{\delta}\right)\Lambda_k e = X + \Lambda_k e$ where X and Λ_k are independent and have respective distributions P_0 (the uniform distribution on M_0) and

the uniform distribution on $[-\Lambda_+, \Lambda_+]$. Therefore, on U_k, $\bar{Q}^{(i)}_{\tau,1}$ has a density with respect to the Lebesgue measure λ_{d+1} on $\mathbb{R}^d \times span(e)$ that can be written as

$$\bar{q}^{(i)}_{\tau,1}(x_1, x_{2:d}, y) = \frac{\mathbb{1}_{[-\Lambda_+,\Lambda_+]}(y)}{2Vol(M_0)\Lambda_+}.$$

Analogously, nearby x_k a random variable Z having distribution $\bar{Q}^{(i)}_{\tau^k,1}$ can be represented as $Z = X + A_k(X - x_k)^i_1 e$ where A_k has uniform distribution on $[A_-, A_+]$. Therefore, a straightforward change of variable yields the density

$$\bar{q}^{(i)}_{\tau^k,1}(x_1, x_{2:d}, y) = \frac{\mathbb{1}_{[A_- x^i_1, A_+ x^i_1]}(y)}{Vol(M_0)\,(A_+ - A_-)\,x^i_1}.$$

We recall that $Vol(M_0) = (2\tau_{min})^d Vol(M^{(0)}_0) = c'_d \tau^d_{min}$. Let us now tackle the right-hand side inequality, writing

$$\begin{aligned}
&\int_{U_k} d\bar{Q}^{(i)}_{\tau,1} \wedge d\bar{Q}^{(i)}_{\tau^k,1} \\
&\quad = \int_{\mathcal{B}(x_k,\delta/2)} \left(\frac{\mathbb{1}_{[-\Lambda_+,\Lambda_+]}(y)}{2Vol(M_0)\Lambda_+} \right) \wedge \left(\frac{\mathbb{1}_{[A_- x^i_1, A_+ x^i_1]}(y)}{Vol(M_0)\,(A_+ - A_-)\,x^i_1} \right) dy dx_1 dx_{2:d} \\
&\quad \geq \int_{\mathcal{B}_{\mathbb{R}^{d-1}}(0,\frac{\delta}{4})} \int_{-\delta/4}^{\delta/4} \int_{\mathbb{R}} \left(\frac{\mathbb{1}_{[-\Lambda_+,\Lambda_+]}(y)}{2\Lambda_+} \right) \wedge \left(\frac{\mathbb{1}_{[A_- x^i_1, A_+ x^i_1]}(y)}{A_+ x^i_1/2} \right) \frac{dy dx_1 dx_{2:d}}{Vol(M_0)}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\int_{U_k} d\bar{Q}^{(i)}_{\tau,1} \wedge d\bar{Q}^{(i)}_{\tau^k,1} \\
&\quad \geq \frac{c_d}{\tau^d_{min}} \delta^{d-1} \int_0^{\delta/4} \int_{A_+ x^i_1/2}^{\Lambda_+ \wedge (A_+ x^i_1)} \frac{1}{2\Lambda_+} \wedge \frac{2}{A_+ x^i_1} dy dx_1 \\
&\quad \geq \frac{c_d}{\tau^d_{min}} \delta^{d-1} \int_0^{\delta/4} \int_{A_+ x^i_1/2}^{(c \wedge 1)(A_+ x^i_1)} \frac{(2c \wedge 1/2)}{2\Lambda_+} dy dx_1 \\
&\quad = \frac{c_d}{\tau^d_{min}} \delta^{d-1} (2c \wedge 1/2)\,(c \wedge 1 - 1/2)\,\frac{A_+}{\Lambda_+} \frac{(\delta/4)^{i+1}}{i+1} \\
&\quad \geq \frac{c_{d,i}}{C} \left(\frac{\delta}{\tau_{min}} \right)^d.
\end{aligned}$$

For the integral on U'_k, notice that by definition, $\bar{Q}^{(i)}_{\tau,n-1}$ and $\bar{Q}^{(i)}_{\tau^k,n-1}$ coincide on U'_k since they are respectively the image distributions of P_0 by functions that are equal on that set. Moreover, these two functions leave $\mathbb{R}^D \setminus \left\{ \mathcal{B}_{\mathbb{R}^d \times \{0\}^{D-d}}(x_k, \delta) + \mathcal{B}_{span(e)}(0, \tau_{min}/2) \right\}$ unchanged. Therefore,

$$\begin{aligned}
&\int_{U'_k} d\bar{Q}^{(i)}_{\tau,n-1} \wedge d\bar{Q}^{(i)}_{\tau^k,n-1} \\
&\qquad = P_0^{\otimes n-1}(U'_k) \\
&\qquad = \left(1 - P_0 \left(\mathcal{B}_{\mathbb{R}^d \times \{0\}^{D-d}}(x_k, \delta) + \mathcal{B}_{span(e)}(0, \tau_{min}/2) \right) \right)^{n-1} \\
&\qquad = \left(1 - \omega_d \delta^d / Vol(M_0) \right)^{n-1},
\end{aligned}$$

hence the result. □

D.3.3 Minimax Inconsistency Results

This section is devoted to the proof of lower bound for tangent space estimation (Theorem VI.6): we build hypotheses P, P' and apply Theorem VI.22. For $\delta \geq \Lambda > 0$, let $\mathcal{C}', \mathcal{C}' \subset \mathbb{R}^3$ be closed curves of the Euclidean space as in Figure VI.1, and such that outside the figure, $\mathcal{C}'$ and $\mathcal{C}'$ coincide and are $\mathcal{C}^\infty$. The bumped parts are obtained with a smooth diffeomorphism similar to (VI.25), centered at x. Here, δ and Λ can be chosen arbitrarily small.

Let $\mathcal{S}^{d-1} \subset \mathbb{R}^d$ be a $d-1$-sphere of radius $1/L_\perp$. Consider the Cartesian products $M_1 = \mathcal{C} \times \mathcal{S}^{d-1}$ and $M_1' = \mathcal{C}' \times \mathcal{S}^{d-1}$. M_1 and M_1' are subsets of $\mathbb{R}^{d+3} \subset \mathbb{R}^D$. Finally, let P_1 and P_1' denote the uniform distributions on M and M'. Note that M, M' can be built by homothecy of ratio $\lambda = 1/L_\perp$ from some unitary scaled $M_1^{(0)}, {M'}_1^{(0)}$, similarly to Section VI.4.2, yielding, from Proposition VI.5, that P_1, P_1' belong to $\mathcal{P}^k_{(x)}$ provided that $L_3/L_\perp^2, \ldots, L_k/L_\perp^{k-1}, L_\perp^d/f_{min}$ and $f_{max}/L_\perp^d$ are large enough (depending only on d and k), and that Λ, δ and Λ^k/δ are small enough. From Le Cam's Lemma VI.22, we have for all $n \geq 1$,

$$\inf_{\hat{T}} \sup_{P \in \mathcal{P}^k_{(x)}} \mathbb{E}_{P^{\otimes n}} \angle(T_x M, \hat{T}) \geq \frac{1}{2} \angle(T_x M_1, T_x M_1') \left(1 - TV(P_1, P_1')\right)^n .$$

By construction, $\angle(T_x M_1, T_x M_1') = 1$, and since $\mathcal{C}$ and $\mathcal{C}'$ coincide outside $\mathcal{B}_{\mathbb{R}^3}(0, \delta)$,

$$\begin{aligned} TV(P_1, P_1') &= Vol\left(\left(\mathcal{B}_{\mathbb{R}^3}(0,\delta) \cap \mathcal{C}\right) \times \mathcal{S}^{d-1}\right) / Vol\left(\mathcal{C} \times \mathcal{S}^{d-1}\right) \\ &= Length\left(\mathcal{B}_{\mathbb{R}^3}(0,\delta) \cap \mathcal{C}\right) / Length(\mathcal{C}) \\ &\leq c_{L_\perp} \delta. \end{aligned}$$

Hence, letting Λ, δ go to 0 with Λ^k/δ small enough, we get the announced bound.

We now tackle the lower bound on second fundamental form estimation with the same strategy. Let $M_2, M_2' \subset \mathbb{R}^D$ be d-dimensional submanifolds as in Figure VI.2: they both contain x, the part on the top of M_2 is a half d-sphere of radius $2/L_\perp$, the bottom part of M_2' is a piece of a d-plane, and the bumped parts are obtained with a smooth diffeomorphism similar to (VI.25) centered at x. Outside $\mathcal{B}(x, \delta)$, M_2, M_2' coincide and connect smoothly the upper and lower parts. Let P_2, P_2' be the probability distributions obtained by the pushforward given by the bump maps. Under the same conditions on the parameters as previously, P_2 and P_2' belong to $\mathcal{P}^k_{(x)}$ according to Proposition VI.5. From Le Cam's Lemma VI.22 we deduce

$$\begin{aligned} \inf_{\widehat{II}} \sup_{P \in \mathcal{P}^k_{(x)}} & \mathbb{E}_{P^{\otimes n}} \left\| II_x^M \circ \pi_{T_x M} - \widehat{II} \right\|_{op} \\ &\geq \frac{1}{2} \left\| II_x^{M_2} \circ \pi_{T_x M_2} - II_x^{M_2'} \circ \pi_{T_x M_2'} \right\|_{op} \left(1 - TV(P_2, P_2')\right)^n . \end{aligned}$$

By construction, $\left\| II_x^{M_2} \circ \pi_{T_x M_2} \right\|_{op} = 0$, and since M_2' is a part of a sphere of radius $2/L_\perp$ nearby x, $\left\| II_x^{M_2'} \circ \pi_{T_x M_2'} \right\|_{op} = L_\perp/2$. Hence,

$$\left\| II_x^{M_2} \circ \pi_{T_x M_2} - II_x^{M_2'} \circ \pi_{T_x M_2'} \right\|_{op} \geq L_\perp/2.$$

Moreover, since P_2 and P_2' coincide on $\mathbb{R}^D \setminus \mathcal{B}(x, \delta)$,

$$TV(P_2, P_2') = Pc_2(\mathcal{B}(x, \delta)) \leq c_{d, L_\perp} \delta^d.$$

Letting Λ, δ go to 0 with Λ^k/δ small enough, we have the desired result.

Bibliography

[AB06a] Stephanie B. Alexander and Richard L. Bishop. Gauss equation and injectivity radii for subspaces in spaces of curvature bounded above. *Geom. Dedicata*, 117:65–84, 2006.

[AB06b] Charalambos D. Aliprantis and Kim C. Border. *Infinite dimensional analysis*. Springer, Berlin, third edition, 2006. A hitchhiker's guide.

[ABE09] Dominique Attali, Jean-Daniel Boissonnat, and Herbert Edelsbrunner. Stability and computation of medial axes: a state-of-the-art report. In *Mathematical foundations of scientific visualization, computer graphics, and massive data exploration*, Math. Vis., pages 109–125. Springer, Berlin, 2009.

[ACLZ17] Ery Arias-Castro, Gilad Lerman, and Teng Zhang. Spectral clustering based on local PCA. *J. Mach. Learn. Res.*, 18:Paper No. 9, 57, 2017.

[ACV14] Ery Arias-Castro and Nicolas Verzelen. Community detection in dense random networks. *Ann. Statist.*, 42(3):940–969, 2014.

[AL13] Dominique Attali and André Lieutier. Optimal reconstruction might be hard. *Discrete Comput. Geom.*, 49(2):133–156, 2013.

[APR16] Ery Arias-Castro, Beatriz Pateiro-López, and Alberto Rodríguez-Casal. Minimax Estimation of the Volume of a Set with Smooth Boundary. *ArXiv e-prints*, May 2016.

[BBL05] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375, 2005.

[BCP08] Gérard Biau, Benoît Cadre, and Bruno Pelletier. Exact rates in density support estimation. *J. Multivariate Anal.*, 99(10):2185–2207, 2008.

[BDWW15] Mickaël Buchet, Tamal K. Dey, Jiayuan Wang, and Yusu Wang. Declutter and resample: Towards parameter free denoising. *ArXiv e-prints*, abs/1511.05479, 2015.

[Bee93] Gerald Beer. *Topologies on closed and closed convex sets*, volume 268 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1993.

[BG14] Jean-Daniel Boissonnat and Arijit Ghosh. Manifold reconstruction using tangential Delaunay complexes. *Discrete Comput. Geom.*, 51(1):221–267, 2014.

[BGO09] 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.

[BH98] Hans U. Bräker and Tailen Hsing. On the area and perimeter of a random convex hull in a bounded convex set. *Probab. Theory Related Fields*, 111(4):517–550, 1998.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[BNS06] 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.

[Bou02] Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

[BRS+12] Sivaraman Balakrishnan, Alessandro Rinaldo, Don Sheehy, Aarti Singh, and Larry A. Wasserman. Minimax rates for homology inference. *Journal of Machine Learning Research - Proceedings Track*, 22:64–72, 2012.

[BRSW13] Sivaraman Balakrishnan, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Tight Lower Bounds for Homology Inference. *ArXiv e-prints*, 2013.

[BSW09] Mikhail Belkin, Jian Sun, and Yusu Wang. Constructing Laplace operator from point clouds in $\mathbb{R}^d$. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1031–1040. SIAM, Philadelphia, PA, 2009.

[BT07] Jean-Daniel Boissonnat and Monique Teillaud, editors. *Effective computational geometry for curves and surfaces*. Mathematics and Visualization. Springer-Verlag, Berlin, 2007.

[BY98] Jean-Daniel Boissonnat and Mariette Yvinec. *Algorithmic geometry*. Cambridge University Press, Cambridge, 1998. Translated from the 1995 French original by Hervé Brönnimann.

[Car09] Gunnar Carlsson. Topology and data. *Bull. Amer. Math. Soc. (N.S.)*, 46(2):255–308, 2009.

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

[CCSL06] Frédéric Chazal, David Cohen-Steiner, and André Lieutier. A sampling theory for compact sets in Euclidean space. In *Computational geometry (SCG'06)*, pages 319–326. ACM, New York, 2006.

[CCSM11] Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for probability measures. *Found. Comput. Math.*, 11(6):733–751, 2011.

[CDR05] Siu-Wing Cheng, Tamal K. Dey, and Edgar A. Ramos. Manifold reconstruction from point samples. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1018–1027. ACM, New York, 2005.

[CFPL12] Antonio Cuevas, Ricardo Fraiman, and Beatriz Pateiro-López. On statistical properties of sets fulfilling rolling-type conditions. *Adv. in Appl. Probab.*, 44(2):311–329, 2012.

[CFRC07] Antonio Cuevas, Ricardo Fraiman, and Alberto Rodríguez-Casal. A nonparametric approach to the estimation of lengths and surface areas. *Ann. Statist.*, 35(3):1031–1051, 2007.

[CGLM15] Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Convergence rates for persistence diagram estimation in topological data analysis. *Journal of Machine Learning Research*, 16:3603–3635, 2015.

[CL05] Frédéric Chazal and André Lieutier. The λ-medial axis. *J. Graphical Models*, 67:304–331, 2005.

[Cla06] Kenneth L. Clarkson. Building triangulations using ε-nets. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 326–335. ACM, 2006.

[CLPL14] Antonio Cuevas, Pamela Llop, and Beatriz Pateiro-López. On the estimation of the medial axis and inner parallel body. *J. Multivariate Anal.*, 129:171–185, 2014.

[CP05] Frédéric Cazals and Marc Pouget. Estimating differential quantities using polynomial fitting of osculating jets. *Comput. Aided Geom. Design*, 22(2):121–146, 2005.

[CRC04] Antonio Cuevas and Alberto Rodríguez-Casal. On boundary estimation. *Adv. in Appl. Probab.*, 36(2):340–354, 2004.

[Cue09] Antonio Cuevas. Set estimation: another bridge between statistics and geometry. *Bol. Estad. Investig. Oper.*, 25(2):71–85, 2009.

[dC92] Manfredo P. do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser Boston, Inc., Boston, MA, 1992. Translated from the second Portuguese edition by Francis Flaherty.

[Dey07] Tamal K. Dey. *Curve and surface reconstruction: algorithms with mathematical analysis*, volume 23 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2007.

[DK70] Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7:1–46, 1970.

[DMGZ94] Giuseppe De Marco, Gianluca Gorni, and Gaetano Zampieri. Global inversion of functions: an introduction. *NoDEA Nonlinear Differential Equations Appl.*, 1(3):229–248, 1994.

[Don95] David L. Donoho. De-noising by soft-thresholding. *IEEE Trans. Inf. Theor.*, 41(3):613–627, May 1995.

[DS06] Tamal K. Dey and Jian Sun. Normal and feature approximations from noisy point clouds. In *FSTTCS 2006: Foundations of software technology and theoretical computer science*, volume 4337 of *Lecture Notes in Comput. Sci.*, pages 21–32. Springer, Berlin, 2006.

[DVW15] Ramsay Dyer, Gert Vegter, and Mathijs Wintraecken. Riemannian simplices and triangulations. *Geom. Dedicata*, 179:91–138, 2015.

[DW96] Lutz Dümbgen and Günther Walther. Rates of convergence for random approximations of convex sets. *Adv. in Appl. Probab.*, 28(2):384–393, 1996.

[Fed59] Herbert Federer. Curvature measures. *Trans. Amer. Math. Soc.*, 93:418–491, 1959.

[Fed69] Herbert Federer. *Geometric measure theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York, 1969.

[FIK+15] Charles Fefferman, Sergei V. Ivanov, Yaroslav Kurylev, Matti Lassas, and Hariharan Narayanan. Reconstruction and interpolation of manifolds I: The geometric Whitney problem. *ArXiv e-prints*, August 2015.

[FLR+14] Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *Ann. Statist.*, 42(6):2301–2339, 2014.

[FMN16] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *J. Amer. Math. Soc.*, 29(4):983–1049, 2016.

[GG12] Jie Gao and Leonidas Guibas. Geometric algorithms for sensor networks. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 370(1958):27–51, 2012.

[GK06] Evarist Giné and Vladimir Koltchinskii. Empirical graph Laplacian approximation of Laplace-Beltrami operators: large sample results. In *High dimensional probability*, volume 51 of *IMS Lecture Notes Monogr. Ser.*, pages 238–259. Inst. Math. Statist., Beachwood, OH, 2006.

[GM11] Michael S. Gashler and Tony Martinez. Tangent space guided intelligent neighbor finding. In *Proceedings of the IEEE International Joint Conference on Neural Networks IJCNN'11*, pages 2617–2624. IEEE Press, 2011.

[GPPVW12a] 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.

[GPPVW12b] Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Minimax manifold estimation. *J. Mach. Learn. Res.*, 13:1263–1291, 2012.

[GSBW11] Xiaoyin Ge, Issam I. Safa, Mikhail Belkin, and Yusu Wang. Data skeletonization via reeb graphs. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 837–845. Curran Associates, Inc., 2011.

[GVL96] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[GWM01] Stefan Gumhold, Xinlong Wang, and Rob MacLeod. Feature Extraction from Point Clouds. In *10th International Meshing Roundtable*, pages 293–305. Sandia National Laboratories,, 2001.

[Har51] Philip Hartman. On geodesic coordinates. *Amer. J. Math.*, 73:949–954, 1951.

[Hat02] Allen Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge, 2002.

[HTF09] 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.

[KA02] Andrew V. Knyazev and Merico E. Argentati. Principal angles between subspaces in an A-based scalar product: algorithms and perturbation estimates. *SIAM J. Sci. Comput.*, 23(6):2008–2040, 2002.

[KMT92] S. Kanagawa, Y. Mochizuki, and H. Tanaka. Limit theorems for the minimum interpoint distance between any pair of i.i.d. random points in $\mathbf{R}^d$. *Ann. Inst. Statist. Math.*, 44(1):121–131, 1992.

[KZ15] 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.

[LC73] Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53, 1973.

[LC98] Erich L. Lehmann and George Casella. *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 1998.

[LV07] John A. Lee and Michel Verleysen. *Nonlinear dimensionality reduction*. Information Science and Statistics. Springer, New York, 2007.

[Mas07] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[Mat75] Georges Matheron. *Random sets and integral geometry*. John Wiley & Sons, New York-London-Sydney, 1975. With a foreword by Geoffrey S. Watson, Wiley Series in Probability and Mathematical Statistics.

[Mey89] Wolfgang Meyer. Toponogov's theorem and its applications. 1989.

[MMS16] 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.

[MOG11] Quentin Mérigot, Maks Ovsjanikov, and Leonidas J. Guibas. Voronoi-based curvature and feature estimation from point clouds. *IEEE Transactions on Visualization and Computer Graphics*, 17(6):743–756, June 2011.

[MS05] Facundo Mémoli and Guillermo Sapiro. Distance functions and geodesics on submanifolds of $\mathbb{R}^d$ and point clouds. *SIAM J. Appl. Math.*, 65(4):1227–1260, 2005.

[MT95] Enno Mammen and Alexander B. Tsybakov. Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.*, 23(2):502–524, 1995.

[Mun75] James R. Munkres. *Topology: a first course*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1975.

[NSW08] 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.

[Oud15] Steve Y. Oudot. *Persistence theory: from quiver representations to data analysis*, volume 209 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2015.

[Par05] Kalyanapuram R. Parthasarathy. *Probability measures on metric spaces*. AMS Chelsea Publishing, Providence, RI, 2005. Reprint of the 1967 original.

[RS00] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.

[Rus04] Szymon Rusinkiewicz. Estimating curvatures and their derivatives on triangle meshes. In *Proceedings of the 3D Data Processing, Visualization, and Transmission, 2Nd International Symposium*, 3DPVT '04, pages 486–493, Washington, DC, USA, 2004. IEEE Computer Society.

[SJ03] Catherine A. Sugar and Gareth M. James. Finding the number of clusters in a dataset. *Journal of the American Statistical Association*, 98(463):750–763, 2003.

[SP07] Alok Sharma and Kuldip K Paliwal. Fast principal component analysis using fixed-point algorithm. *Pattern Recognition Letters*, 28(10):1151–1155, 2007.

[SW12] Amit Singer and Hau-tieng Wu. Vector diffusion maps and the connection Laplacian. *Comm. Pure Appl. Math.*, 65(8):1067–1144, 2012.

[TdSL00] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319, 2000.

[Thä08] Christoph Thäle. 50 years sets with positive reach—a survey. *Surv. Math. Appl.*, 3:123–165, 2008.

[Tod79] Isaac Todhunter. *Spherical Trigonometry, for the Use of Colleges and Schools: With Numerous Examples*. Macmillan, 1879.

[Tsy09] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[TVF13] Hemant Tyagi, Elif Vural, and Pascal Frossard. Tangent space estimation for smooth embeddings of Riemannian manifolds. *Inf. Inference*, 2(1):69–114, 2013.

[UM14] Konstantin Usevich and Ivan Markovsky. Optimization on a grassmann manifold with application to system identification. *Automatica*, 50(6):1656 – 1662, 2014.

[Was] Larry Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5(1).

[Yu97] Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Vitesses de convergence en inférence géométrique

Mots clefs : Statistiques non-asymptotiques, inférence géométrique, vitesses minimax, apprentissage de variétés.

Résumé : Certains jeux de données présentent des caractéristiques géométriques et topologiques non triviales qu'il peut être intéressant d'inférer. Cette thèse traite des vitesses non-asymptotiques d'estimation de différentes quantités géométriques associées à des sous-variétés $M \subset \mathbb{R}^D$. Dans chaque cas, on dispose d'un n-échantillon i.i.d. de loi commune P ayant pour support M. On étudie le problème d'estimation de la sous-variété M pour la perte donnée par la distance de Hausdorff, du reach τ_M, de l'espace tangent $T_X M$ et de la seconde forme fondamentale II_X^M, pour $X \in M$ à la fois déterministe et aléatoire. Les vitesses sont données en fonction la taille n de l'échantillon, de la dimension intrinsèque de M ainsi que de sa régularité. Dans l'analyse, on obtient des résultats de stabilité pour des techniques de reconstruction existantes, une procédure de débruitage ainsi que des résultats sur la géométrie du reach τ_M. Une extension du lemme d'Assouad est exposée, permettant l'obtention de bornes inférieures minimax dans des cadres singuliers.

Title: Rates of Convergence for Geometric Inference

Keys words: Non-asymptotic statistics, geometric inference, minimax rates, manifold learning

Abstract: Some datasets exhibit non-trivial geometric or topological features that can be interesting to infer. This thesis deals with non-asymptotic rates for various geometric quantities associated with submanifolds $M \subset \mathbb{R}^D$. In all the settings, we are given an i.i.d. n-sample with common distribution P having support M. We study the optimal rates of estimation of the submanifold M for the loss given by the Hausdorff metric, of the reach τ_M, of the tangent space $T_X M$ and the second fundamental form II_X^M, for $X \in M$ both deterministic and random. The rates are given in terms of the sample size n, the instrinsic dimension of M and of its regularity. In the process, we obtain stability results for existing reconstruction techniques, a denoising procedure and results on the geometry of the reach τ_M. An extension of Assouad's lemma is presented, allowing to derive minimax lower bounds in singular frameworks.