Approximation and Geometry of the Reach

Eddie Aamari

Inria Saclay, Université d'Orsay

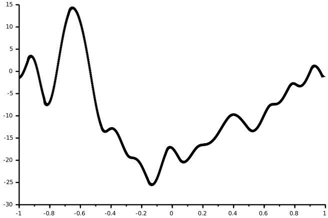
Gudhi/TopData Workshop 2016, Porquerolles

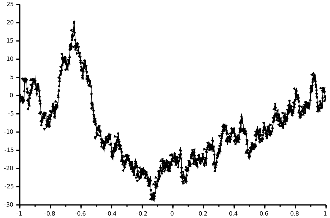
10/20/2016

With J. Kim, F. Chazal, B. Michel, A. Rinaldo, L. Wasserman

Regularity

Regularity and **scale** parameters are crucial in approximations problems, and in actual implementation for estimation.





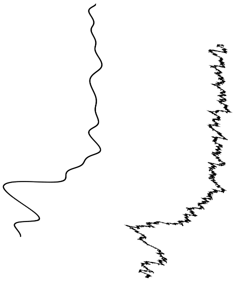
Classical regularity classes: Hölder, Sobolev, Besov, ...?
Such classes allow to control variations in the form of increments

$$\|f(x) - f(y)\| \leq K \|x - y\|^{\alpha}.$$

$\rightarrow$ Drives the difficulty of the statistical problem.

Regularity Without Coordinates?

Without natural coordinates, usual increments "$\|f(x) - f(y)\|$" no longer make sense.
Need for an intrinsic way to describe the difficulty of a problem.



Some computational geometers and statisticians use the **reach**.

Bibliography

First introduced by H. Federer (1957), the reach is a regularity and scale parameter that has recently grown popular in the geometric inference literature.

- **Homology Inference:** Niyogi, Smale, Weinberger, Dey, Lieutier
- **Manifold Reconstruction:** Boissonnat, Ghosh, CMU TopStat group
- **Volume Estimation:** Cuevas, Fraiman, Pateiro-López, Rodríguez-Casal
- **Manifold Clustering:** Arias-Castro, Lerman, Zhang

Medial Axis

The **medial axis** of $M \subset \mathbb{R}^D$ is the set of points that have at least two nearest neighbors on M.

$$Med(M) = \{z \in \mathbb{R}^D, z \text{ has several nearest neighbors on } M\},$$

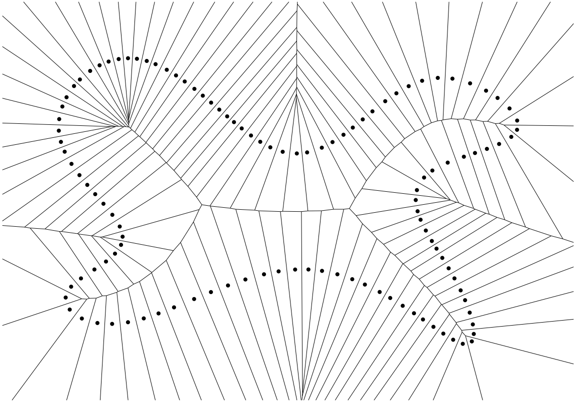


Figure : Voronoi diagram of a point cloud

Medial Axis

The **medial axis** of $M \subset \mathbb{R}^D$ is the set of points that have at least two nearest neighbors on M.

$$Med(M) = \{z \in \mathbb{R}^D, z \text{ has several nearest neighbors on } M\},$$

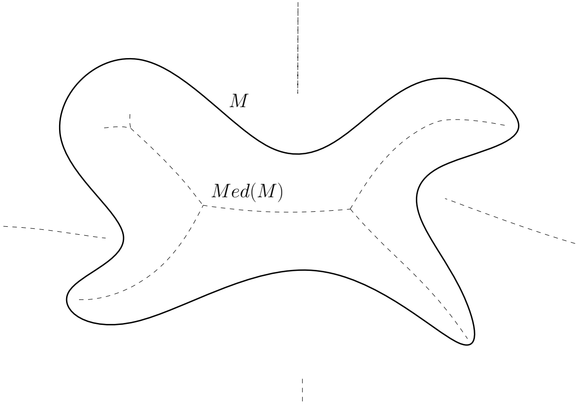


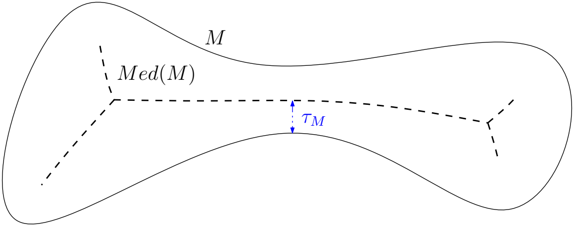
Figure : Medial axis of a continuous subset

Reach

For a closed subset $M \subset \mathbb{R}^D$, the **reach** τ_M of M is the least distance to its medial axis.

$$\tau_M = \inf_{x \in M} d\left(x, med(M)\right),$$

where $d(x, A) = \inf_{a \in A} \|x - a\|$ for all $x \in \mathbb{R}^D$.



One can also flip the formula, in the sense that

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

Global Regularity

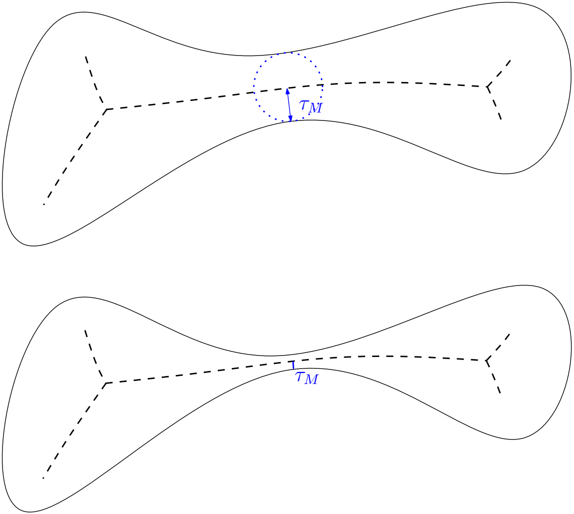


Figure : The smaller τ_M, the tighter a bottleneck structure is possible.

Local Regularity

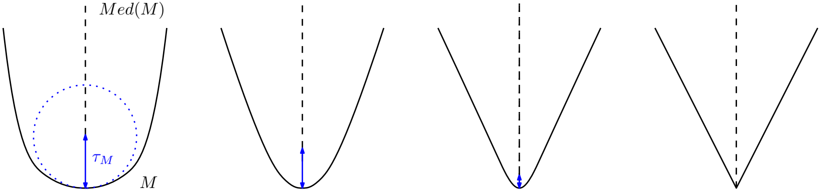


Figure : High curvature $\equiv$ Small radius of curvature $\equiv \tau_M \to 0$.

Proposition (Nyiogi, Smale, Weinberger — 2006)

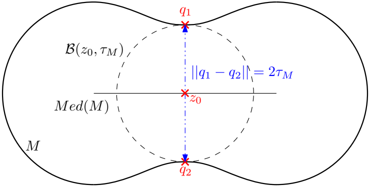
Let II denote the second fundamental form of M. For all unit tangent vector $v \in T_x M$, $II_x(v, v) \leq 1/\tau_M$.

Proposition (Dey, Li — 2009)

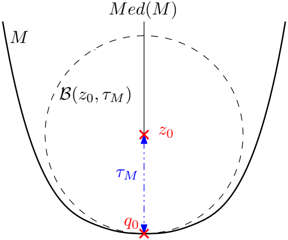
The sectional curvatures κ satisfy $|\kappa| \leq 2/\tau_M^2$.

Theorem (A,K,C,M,R,W — 2016?)

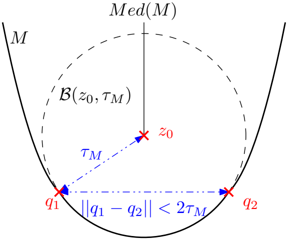
For a closed $\mathcal{C}^3$ submanifold $M \subset \mathbb{R}^D$, the reach can be attained with:



- *A bottleneck.*



- *No reach attaining pair.*



- *A reach attaining pair,*
- *No bottleneck.*

Global and Local Reach

Corollary

Let $M \subset \mathbb{R}^D$ be a closed $\mathcal{C}^3$ submanifold with reach τ_M. At least one of the following two assertions holds.

- ***(Global case)*** *M has a bottleneck $(q_1, q_2) \in M^2$, i.e. 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 $\gamma_0 = \gamma_{q_0,v_0}$ such that $\gamma_0(0) = q_0$ and $\|\gamma_0''(0)\| = \frac{1}{\tau_M}$.*

Global and Local Reach

Corollary

Let $M \subset \mathbb{R}^D$ be a closed $\mathcal{C}^3$ submanifold with reach τ_M. At least one of the following two assertions holds.

- ***(Global case)*** *M has a bottleneck $(q_1, q_2) \in M^2$, i.e. 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 $\gamma_0 = \gamma_{q_0,v_0}$ such that $\gamma_0(0) = q_0$ and $\|\gamma_0''(0)\| = \frac{1}{\tau_M}$.*

In view of estimation: we do not know a priori which case the underlying M belongs to.
$\rightarrow$ An estimator should handle both cases indiscriminately.

Geometric and Statistical Model

Definition (Geometric Model)

We let $\mathcal{M}^{d,D}_{\tau_{min},L}$ *denote the set of connected compact submanifolds* $M \subset \mathbb{R}^D$ *without boundary, such that* $\tau_M \geq \tau_{min} > 0$*, and for which every arc-length parametrized geodesic* $\gamma_{p,v}$ *is* $\mathcal{C}^3$ *and satisfies*

$$\left\|\gamma'''_{p,v}(0)\right\| \leq L.$$

Definition (Statistical Model)

We let $\mathcal{Q}^{d,D}_{\tau_{min},L,f_{min}}$ *denote the set of distributions* Q *having support* $M \in \mathcal{M}^{d,D}_{\tau_{min},L}$ *and with a density* $f = \frac{\mathrm{d}Q}{\mathrm{d}vol_M} \geq f_{min} > 0$ *on* M*.*

From now on, we assume that **the tangent spaces are known** at observed points. Data takes the form $(X_1, T_{X_1}M), \ldots, (X_n, T_{X_n}M)$.

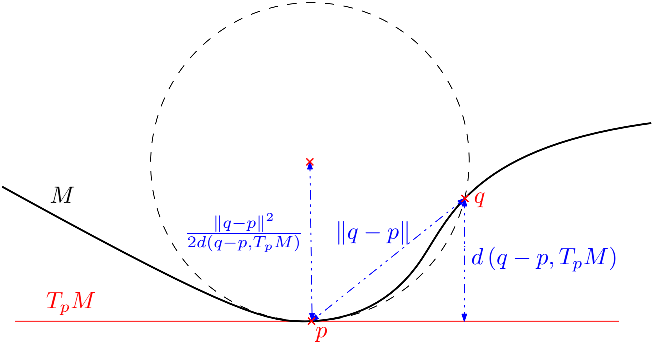
In these models, estimating τ_M is equivalent to estimate $1/\tau_M$.

A (Crucial) Local Formulation

Proposition (Federer — 1957)

For all closed submanifold $M \subset \mathbb{R}^D$,

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



A (Crucial) Local Formulation

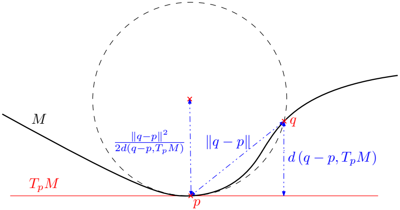
Proposition (Federer — 1957)

For all closed submanifold $M \subset \mathbb{R}^D$,

$$\tau_M = \inf_{p \neq q \in M} \frac{\|q-p\|^2}{2d\left(q-p, T_p M\right)}.$$

Plugin Estimator: Let $\mathbb{X} = \{x_1, \ldots, x_n\} \subset M$ be a finite point cloud. Define

$$\hat{\tau}\left(\mathbb{X}\right) = \inf_{x_i \neq x_j \in \mathbb{X}} \frac{\|x_j - x_i\|^2}{2d(x_j - x_i, T_{x_i} M)}.$$



A (Crucial) Local Formulation

Proposition (Federer — 1957)

For all closed submanifold $M \subset \mathbb{R}^D$,

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

Plugin Estimator: Let $\mathbb{X} = \{x_1, \ldots, x_n\} \subset M$ be a finite point cloud. Define

$$\hat{\tau}\,(\mathbb{X}) = \inf_{x_i \neq x_j \in \mathbb{X}} \frac{\|x_j - x_i\|^2}{2d(x_j - x_i,\, T_{x_i} M)}.$$

$\hat{\tau}$ is decreasing for inclusion: if $\mathbb{Y} \subset \mathbb{X} \subset M$,

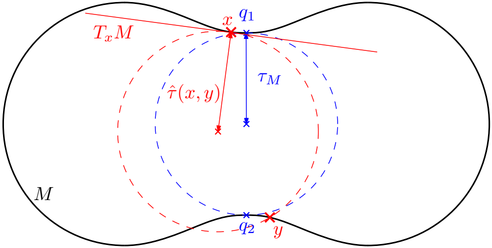
$$\hat{\tau}(\mathbb{Y}) \geq \hat{\tau}(\mathbb{X}) \geq \hat{\tau}(M) = \tau_M.$$

Global Case

Proposition (A,K,C,M,R,W — 2016?)

Let $M \subset \mathbb{R}^D$ be a submanifold with reach τ_M that has a bottleneck $q_1, q_2 \in M$. Let $\mathbb{X} \subset M$.
If there exist $x, y \in \mathbb{X}$ with $\|q_1 - x\| < \tau_M$ and $\|q_2 - y\| < \tau_M$,

$$\frac{1}{\tau_M} \geq \frac{1}{\hat{\tau}(\mathbb{X})} \geq \frac{1}{\hat{\tau}(\{x, y\})} \geq \frac{1}{\tau_M} - \frac{9}{2\tau_M^2} \max\{d_M(q_1, x), d_M(q_2, y)\}.$$



Minimax Estimate in the Global Case

If $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ is a i.i.d. sample, the integrated bound follows by lower bounding the probability to get two points X_i and X_j close to q_1 and q_2.

Corollary

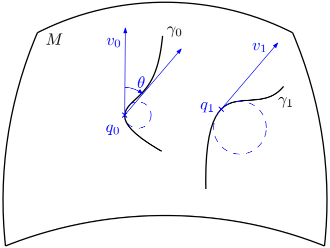
Let $P \in \mathcal{P}^{d,D}_{\tau_{\min},L,f_{\min}}$ and $M = supp(P)$. Assume M has a bottleneck $q_1, q_2 \in M$. Then,

$$\mathbb{E}_{P^n}\left[\left|\frac{1}{\hat{\tau}(\mathbb{X}_n)} - \frac{1}{\tau_M}\right|^p\right] \leq C_{p,d,\tau_{min},L,f_{min}} n^{-\frac{p}{d}},$$

where $C_{p,d,\tau_{\min},L,f_{\min}}$ depends only on p, d, τ_{min},L and $f_{\min}$.

Local Case

Assume there exist $q_0 \in M$ and $v_0 \in T_{q_0}M$ with $\left\|\gamma''_{q_0,v_0}(0)\right\| = 1/\tau_M$.

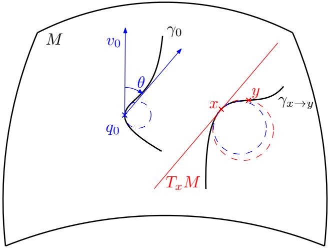


- (Principal Curvature Stability). If $d_M(q_0, q_1)$ and $\theta = \angle(v_0, v_1)$ are small,

$$\left\|\gamma''_{q_1,v_1}(0)\right\| \simeq \left\|\gamma''_{q_0,v_0}(0)\right\| = 1/\tau_M.$$

Local Case

Assume there exist $q_0 \in M$ and $v_0 \in T_{q_0}M$ with $\left\|\gamma''_{q_0,v_0}(0)\right\| = 1/\tau_M$.



- (Principal Curvature Stability). If $d_M(q_0, q_1)$ and $\theta = \angle(v_0, v_1)$ are small,

$$\left\|\gamma''_{q_1,v_1}(0)\right\| \simeq \left\|\gamma''_{q_0,v_0}(0)\right\| = 1/\tau_M.$$

- (Directional Curvature Estimation). Write $\gamma_{x \to y}$ for the geodesic joining x to y. If $\|y - x\|$ is small,

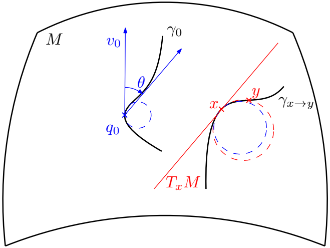
$$\frac{\|y - x\|^2}{2d\left(y - x,\, T_xM\right)} \simeq \left\|\gamma''_{x \to y}(0)\right\|.$$

Local Case

Proposition (A,K,C,M,R,W — 2016?)

Let $M \in \mathcal{M}^{d,D}_{\tau_{\min},L}$ be such that there exist $q_0 \in M$ and a geodesic γ_0 with $\gamma_0(0) = q_0$ and $\|\gamma_0''(0)\| = \frac{1}{\tau_M}$.
Let $\mathbb{X} \subset M$ and $x, y \in \mathbb{X}$ be such that $x, y \in \mathcal{B}_M\left(q_0, \frac{\tau_M}{4}\right)$. Let $\gamma_{x\to y}$ be the geodesic joining x and y and $\theta = \angle\left(\gamma_0'(0), \gamma_{x\to y}'(0)\right)$.

$$\frac{1}{\tau_M} \geq \frac{1}{\hat{\tau}(\mathbb{X})} \geq \frac{1}{\hat{\tau}(\{x,y\})} \geq \frac{1}{\tau_M} - \left\{ \frac{4\sin^2\theta}{\tau_M} + \frac{37 d_M(x,y)^2}{\tau_M^3} + \left(\frac{8}{\tau_M^3} + L\right) d_M(x,y) + \frac{2}{3} L d_M(q_0, x) \right\}.$$



Minimax Estimate in the Local Case

If $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ is a i.i.d. sample, the integrated bound follows by lower bounding the probability to get two points X_i, X_j close to q_0 and almost aligned with v_0.

Corollary (A,K,C,M,R,W — 2016?)

Let $P \in \mathcal{P}^{d,D}_{\tau_{\min},L,f_{\min}}$ and $M = supp(P)$. Suppose there exists $q_0 \in M$ and a geodesic γ_0 such that $\gamma_0(0) = q_0$ and $\|\gamma_0''(0)\| = \frac{1}{\tau_M}$. Then,

$$\mathbb{E}_{P^n}\left[\left|\frac{1}{\hat{\tau}(\mathbb{X}_n)} - \frac{1}{\tau_M}\right|^p\right] \leq C_{\tau_{\min},L,f_{\min}}\, n^{-\frac{4p}{5d-1}},$$

where $C_{\tau_{\min},L,f_{\min}}$ depends only on $\tau_{\min}$, L and $f_{\min}$.

Minimax Risk

Let us denote by R_n the minimax risk over $\mathcal{P}^{d,D}_{\tau_{min},L,f_{min}}$.

$$R_n^p = \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}^{d,D}_{\tau_{min},L,f_{min}}} \mathbb{E}_{P^n} \left| \frac{1}{\tau_P} - \frac{1}{\hat{\tau}_n} \right|^p,$$

where the infimum is taken over all the estimators $\hat{\tau}_n$ computed over an n-sample $(X_1, T_{X_1}), \ldots, (X_n, T_{X_n})$.

Minimax Risk

Let us denote by R_n the minimax risk over $\mathcal{P}^{d,D}_{\tau_{min},L,f_{min}}$.

$$R_n^p = \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}^{d,D}_{\tau_{min},L,f_{min}}} \mathbb{E}_{P^n} \left| \frac{1}{\tau_P} - \frac{1}{\hat{\tau}_n} \right|^p,$$

where the infimum is taken over all the estimators $\hat{\tau}_n$ computed over an n-sample $(X_1, T_{X_1}), \ldots, (X_n, T_{X_n})$.

Corollary

For n large enough,

$$R_n^p \leq C_{p,\tau_{\min},L,f_{\min}} n^{-\frac{4p}{5d-1}},$$

for some constant $C_{p,\tau_{\min},L,f_{\min}}$ depending only on $p, \tau_{\min}, L$ and $f_{\min}$.

Minimax Risk

Let us denote by R_n the minimax risk over $\mathcal{P}^{d,D}_{\tau_{min},L,f_{min}}$.

$$R_n^p = \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}^{d,D}_{\tau_{min},L,f_{min}}} \mathbb{E}_{P^n} \left| \frac{1}{\tau_P} - \frac{1}{\hat{\tau}_n} \right|^p,$$

where the infimum is taken over all the estimators $\hat{\tau}_n$ computed over an n-sample $(X_1, T_{X_1}), \ldots, (X_n, T_{X_n})$.

Proposition (A,K,C,M,R,W — 2016?)

Assume that $(4\pi)^d \tau_{min}^d \leq f_{min}^{-1}/2$, $L \geq \frac{1}{2\tau_{min}^2}$ *and* $D \geq 2d$. *Then,*

$$c_{p,\tau_{min}} n^{-p/d} \leq R_n^p \leq C_{p,\tau_{\min},L,f_{\min}} n^{-\frac{4p}{5d-1}},$$

for n *large enough.*

Le Cam's Lemma

For two probability distributions Q, Q' on $\mathbb{R}^D$, the **total variation** distance between them is

$$TV(Q, Q') = \sup_{B \in \mathcal{B}(\mathbb{R}^D)} |Q(B) - Q'(B)| .$$

Theorem (L. Le Cam)

Let $Q, Q' \in \mathcal{Q}^{d,D}_{\tau_{min}, L, f_{min}}$ with respective supports M and M'. Then for all $n \geq 1$,

$$R_n^p \geq c_p \left| \frac{1}{\tau_M} - \frac{1}{\tau_{M'}} \right|^p (1 - TV(Q, Q'))^{2n} .$$

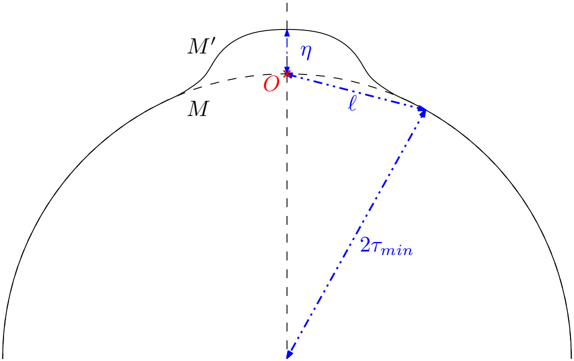
Deriving a minimax lower bound amounts to find Q, Q' such that:

- $\left| \frac{1}{\tau_M} - \frac{1}{\tau_{M'}} \right|$ is large,
- $TV(Q, Q')$ is small.

Le Cam's Lemma Heuristic

For $\eta \approx \ell^3$ and $\ell^d \approx 1/n$,

- $\left|\frac{1}{\tau_M} - \frac{1}{\tau_{M'}}\right| \gtrsim \left(\frac{1}{n}\right)^{1/d}$,
- with high probability, a n-sample does not separate M and M'.



What if Tangent Spaces are Unknown?

Given a point cloud $\mathbb{X} \subset \mathbb{R}^D$ and a family $T = \{T_x\}_{x \in \mathbb{X}}$ of linear subspaces of $\mathbb{R}^D$ indexed by $\mathbb{X}$, the plug-in estimator is defined as

$$\hat{\tau}(\mathbb{X}, T) = \inf_{x \neq y \in \mathbb{X}} \frac{\|y - x\|^2}{2d(y - x, T_x)}.$$

This generalises the previous estimator $\hat{\tau}(\mathbb{X}) = \hat{\tau}(\mathbb{X}, TM)$. Notice that,

$$\tau_M = \inf_{x \neq y \in M} \frac{\|y - x\|^2}{2d(y - x, T_x M)} = \hat{\tau}(M, TM).$$

Tangent Space Stability

For two linear subspaces $U, V \in \mathbb{G}^{d,D}$, let $\angle(U, V) = \|\pi_U - \pi_V\|_{op}$ denote their principal angle.

Proposition

Let $\mathbb{X}$ be a subset of $\mathbb{R}^D$ and $T = \{T_x\}_{x \in \mathbb{X}}$, $\tilde{T} = \{\tilde{T}_x\}_{x \in \mathbb{X}}$ be two families of linear subspaces of $\mathbb{R}^D$ indexed by $\mathbb{X}$.
Assume $\mathbb{X}$ to be δ-sparse, T and $\tilde{T}$ to be θ-close, in the sense that

$$\inf_{x \neq y \in \mathbb{X}} \|y - x\| \geq \delta \quad \textit{and} \quad \sup_{x \in \mathbb{X}} \angle(T_x, \tilde{T}_x) \leq \theta.$$

Then,

$$\left| \frac{1}{\hat{\tau}(\mathbb{X}, T)} - \frac{1}{\hat{\tau}(\mathbb{X}, \tilde{T})} \right| \leq \frac{2\theta}{\delta}.$$

Corollary

All the previous deterministic upper bounds hold for $\hat{\tau}\left(\mathbb{X}, \tilde{T}\right)$ with an extra error term $2\theta/\delta$.

Yet to Be Done

- Finish to write the paper...
- Make the minimax upper and lower bounds match.
- Include noise. For this, it could boil down to prove that the model $\mathcal{M}^{d,D}_{\tau_{min},L}$ is stable under the action of $\mathcal{C}^3$-diffeomorphisms.
- Give minimax upper bounds with unknown tangent spaces.
- Tackle related regularity parameters such as λ-reach, μ-reach or local feature size.

Yet to Be Done

- Finish to write the paper...
- Make the minimax upper and lower bounds match.
- Include noise. For this, it could boil down to prove that the model $\mathcal{M}^{d,D}_{\tau_{min},L}$ is stable under the action of $\mathcal{C}^3$-diffeomorphisms.
- Give minimax upper bounds with unknown tangent spaces.
- Tackle related regularity parameters such as λ-reach, μ-reach or local feature size.

Thanks