Dimension reduction and manifold learning

Dimension estimation, manifold estimation

Eddie Aamari
Département de mathématiques et applications
CNRS, ENS PSL

Master MASH — Dauphine PSL

Dimension estimation

Hyperparameters

Most of the methods crucially required two types of parameters.

Bandwidths

- For building neighborhood graphs (All methods)
  - k-nearest neighbors
  - r-neighborhood
- For building functions on the graph
  - Kernel scale σ or t (Laplacian methods, k-PCA)
  - Localization radius h in local PCA (Hessian LLE, LTSA)

Dimension

- For determining the output dimension

Dimension is inherent to *dimensionality reduction*.

Intrinsic Dimension

Pioneers in intrinsic dimension estimation

This question dates back to **bennett1969intrinsic** in signal processing.

> The intrinsic dimensionality of a collection of signals is defined to be equal to the number of free parameters required in a hypothetical signal generator capable of producing a close approximation to each signal in the collection. Thus defined, the dimensionality becomes a relationship between the vectors representing the signals.

Overview

See **camastra2016intrinsic** for a recent survey.

A Statistical Remark

We are trying to estimate a discrete quantity

$$d \in \{0, \ldots, p\}$$

If $x_1, \ldots, x_n \sim_{iid} \mathrm{Unif}_M$, we hence expect fast estimation rates of the form

$$\mathbb{P}\big(\hat{d} \neq \dim(M)\big) \leq C \exp(-C'n),$$

where $\hat{d} = \hat{d}(x_1, \ldots, x_n)$ is some wisely chosen estimator.

A Statistical Remark

We are trying to estimate a discrete quantity

$$d \in \{0, \dots, p\}$$

If $x_1, \dots, x_n \sim_{iid} \mathrm{Unif}_M$, we hence expect fast estimation rates of the form

$$\mathbb{P}\big(\hat{d} \neq \dim(M)\big) \leq C \exp(-C'n),$$

where $\hat{d} = \hat{d}(x_1, \dots, x_n)$ is some wisely chosen estimator.

A Take-Away Message

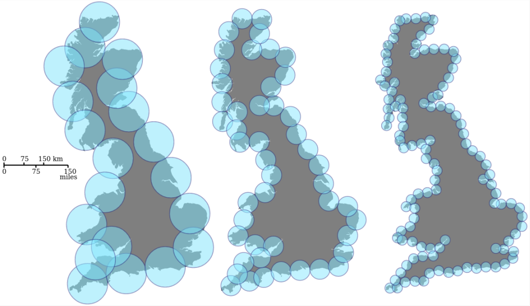
$$\# \{\text{Definitions of dimension}\} \asymp \# \{\text{Estimators of dimension}\} \gg 1$$

Hausdorff and Information Dimension

The Hausdorff dimension $\dim_H(M)$ of $M \subset \mathbb{R}^p$ is defined through

$$\Gamma_H^{(d)} := \lim_{r \to 0^+} \inf_{\substack{x_1,\dots,x_N \in S \\ r_i \le r \\ \cup_i B(x_i, r_i) \supset M}} \sum_i r_i^d,$$

and $\dim_H(M) := \inf \left\{ d \mid \Gamma_H^{(d)} = 0 \right\} \in [0, p]$

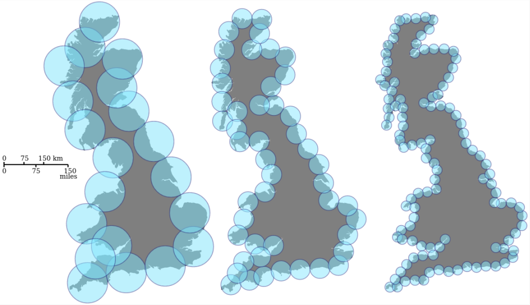


Hausdorff and Information Dimension

Definition

The information dimension $\dim_H(P)$ of a probability measure P, is the smallest Hausdorff dimension of sets that have measure 1.

(For non-pathological cases, $\dim_H(P) = \dim_H(\text{Support}(P))$)



(Generalized) Traveling Salesman Problem

kim2019minimax study the testing problem

$$\mathcal{H}_0\colon \dim(M) = d_0 \qquad \text{VS} \qquad \mathcal{H}_1\colon \dim(M) = d_1$$

where $1 \leq d_0 < d_1 \leq p$ are fixed.

Generalized TSP Leverage of the behavior of the generalized Travelling Salesman Problem (TSP) value

$$\mathrm{TSP}_{d_0}(\mathcal{X}) := \min_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{n} \left\| x_{\sigma(i+1)} - x_{\sigma(i)} \right\|^{d_0},$$

where $\mathfrak{S}_n$ is the set of permutations of $\{1, \ldots, n\}$.

(Generalized) Traveling Salesman Problem

kim2019minimax study the testing problem

$$\mathcal{H}_0\colon \dim(M) = d_0 \qquad \text{VS} \qquad \mathcal{H}_1\colon \dim(M) = d_1$$

where $1 \leq d_0 < d_1 \leq p$ are fixed.

Generalized TSP Leverage of the behavior of the generalized Travelling Salesman Problem (TSP) value

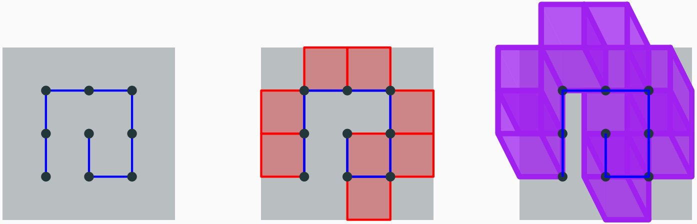
$$\mathrm{TSP}_{d_0}(\mathcal{X}) := \min_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{n} \|x_{\sigma(i+1)} - x_{\sigma(i)}\|^{d_0},$$

where $\mathfrak{S}_n$ is the set of permutations of $\{1, \ldots, n\}$.

Intractability
Generalized TSP is NP-complete

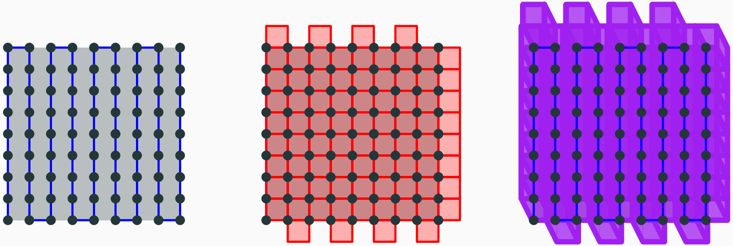
Insights Behind Generalized TSP

$$\mathrm{TSP}_{d_0}(\mathcal{X}) := \min_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{n} \left\| x_{\sigma(i+1)} - x_{\sigma(i)} \right\|^{d_0}$$



Insights Behind Generalized TSP

$$\mathrm{TSP}_{d_0}(\mathcal{X}) := \min_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{n} \|x_{\sigma(i+1)} - x_{\sigma(i)}\|^{d_0}$$



$d_0 < \dim(M)$
$\mathrm{TSP}_{d_0}(\mathcal{X}) \to \infty$

$d_0 = \dim(M)$
$\mathrm{TSP}_{d_0}(\mathcal{X}) = \Theta(1)$

$d_0 > \dim(M)$
$\mathrm{TSP}_{d_0}(\mathcal{X}) \to 0$

TSP Test

For $1 \leq d_0 < d_1 \leq p$ fixed,

$$\hat{d}(\mathcal{X}) := \begin{cases} d_0 & \text{if } \mathrm{TSP}_{d_0}(\mathcal{X}) \leq C, \\ d_1 & \text{otherwise.} \end{cases}$$

Convergence Result

Theorem (kim2019minimax)

Assume that M is $\mathcal{C}^2$ smooth and $x_1, \ldots, x_n \sim_{iid} P$ uniform on M. If $\dim(M) \in \{d_0, d_1\}$, then

$$\mathbb{P}\big(\hat{d}(\mathcal{X}) \neq \dim(M)\big) \lesssim 1_{\dim(M)=d_1} \left(\frac{1}{n}\right)^{\left(\frac{d_1}{d_0}-1\right)n}$$

TSP Estimator

Define

$$\hat{d}(\mathcal{X}) := \min \{ d_0 \, | \mathrm{TSP}_{d_0}(\mathcal{X}) \leq C_{d_0} \}$$

Convergence Result

Theorem (kim2019minimax)

If that M is $\mathcal{C}^2$ smooth and $x_1, \ldots, x_n \sim_{iid} P$ uniform on M, then

$$\mathbb{P}\big(\hat{d}(\mathcal{X}) \neq \dim(M)\big) \lesssim \left(\frac{1}{n}\right)^{\frac{n}{p-1}}$$

Differential / Topological Dimension

Definition

The topological dimension of $M \subset \mathbb{R}^p$ is *the* dimension $\dim_R(M)$ of the model space that locally parametrizes it.

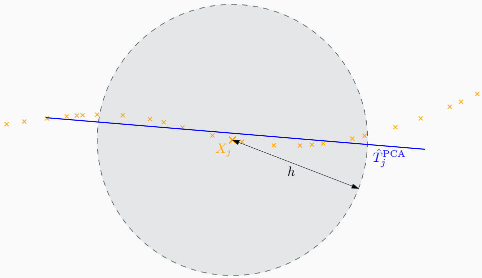
$\Rightarrow$ Local flatness (Essentially assuming manifold structure)

Differential / Topological Dimension

Definition

The topological dimension of $M \subset \mathbb{R}^p$ is *the* dimension $\dim_R(M)$ of the model space that locally parametrizes it.

$\Rightarrow$ Local flatness (Essentially assuming manifold structure)



$\hookrightarrow$ Thresholding principal components **fukunaga1971algorithm**

Local linear residuals

Step 1: Localize

Pick a point $x \in \mathcal{X}$, a localization radius $h > 0$, and set

$$\tilde{\mathcal{X}}_h(x) := \mathcal{X} \cap B(x, h) - x$$

Step 2: Singular Value Decomposition

Compute the SVD of the matrix associated with $\tilde{\mathcal{X}}_h(x)$. Store the singular values $\lambda_1 \geq \ldots \geq \lambda_p$.

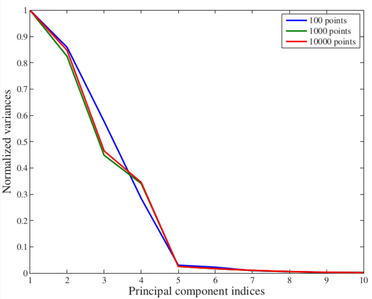
Step 3: Vary thresholds

Plot the residual error (or explained variance)

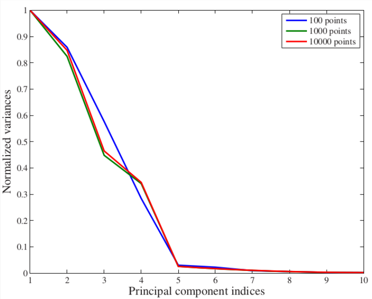
$$d \mapsto \frac{\sum_{k=d+1}^{p} \lambda_k}{\sum_{k=1}^{p} \lambda_k}$$

and search for a gap.

Illustration



Illustration



From Local to Global

In practice, need to aggregate the estimated dimensions $\hat{d}(x)$.

Trial and Error

Such a post-hoc error measurement also applies to any (local) MDS-based dimension reduction technique.

Nearest Neighbors

Instead of fixing a bandwidth, one can also regressing k-Nearest Neighbor distances

fukunaga1971algorithm show that if $x_1, \ldots, x_n \sim_{iid} f(x)\lambda_d(\mathrm{d}x)$ and $x \in \mathbb{R}^d$ with $f(x) > 0$ continuous at x, then

$$\mathbb{E}[\|x_{(k)} - x\|] \propto k^{1/d}$$

Nearest Neighbors

Instead of fixing a bandwidth, one can also regressing k-Nearest Neighbor distances

fukunaga1971algorithm show that if $x_1, \ldots, x_n \sim_{iid} f(x)\lambda_d(\mathrm{d}x)$ and $x \in \mathbb{R}^d$ with $f(x) > 0$ continuous at x, then

$$\mathbb{E}[\|x_{(k)} - x\|] \propto k^{1/d}$$

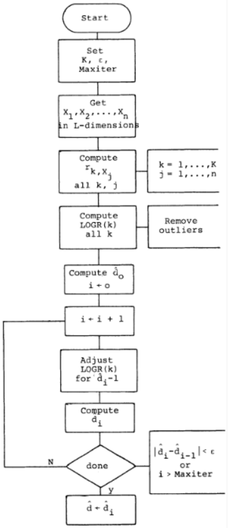


Fig. 1. Procedure for computing $\hat{d}$.

Correlation Dimension

Leveraging local neighborhood properties can also be done by noticing that if $x, x' \overset{\perp\!\!\!\perp}{\sim} \lambda_d$,

$$\mathbb{P}(\|x - x'\| \leq r) \asymp r^d.$$

This leads to the correlation dimension, based on

$$\mathrm{Cor}_r^{(2)}(P) := \mathbb{P}_{x,x' \overset{\perp\!\!\!\perp}{\sim} P}(\|x - x'\| \leq r),$$

and defined as

$$\dim_{\mathrm{cor},2}(P) := \lim_{r \to 0} \frac{\log \mathrm{Cor}_r^{(2)}(P)}{\log r}$$

Correlation Dimension

Estimator

Given $x_1, \ldots, x_n \sim_{iid} P$, consider the U-statistic

$$\widehat{\mathrm{Cor}}_r^{(2)} := \frac{2}{n(n-1)} \sum_{i<j} \mathbf{1}_{\|x_i - x_j\| \leq r},$$

with associate dimension estimator

$$\hat{d}_{\mathrm{cor},2} := \lim_{r \to 0} \frac{\log \widehat{\mathrm{Cor}}_r^{(2)}}{\log r}$$

Convergence

See results in **kegl2002intrinsic** and **higher-order**.

Covering Number

Definition

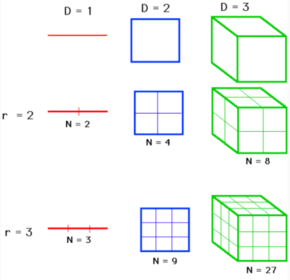
Given $M \subset \mathbb{R}^p$, the r-covering number of M is

$$\mathrm{cv}_M(r) := \min\left\{N \,\middle|\, \exists z_1, \ldots, z_N \in \mathbb{R}^p \text{ s.t. } M \subset \cup_{i=1}^N B(x_i, r)\right\}$$

The r-dimension of M is $\dim_r(M) := \dfrac{\log \mathrm{cv}_M(r)}{-\log r}$.

For $M = [0,1]^d$,

$$\mathrm{cv}_M(r) \asymp \left(\frac{1}{r}\right)^d$$



Minkowski / Capacity Dimension

Definition

The Minkowski (or Capacity) dimension of M is

$$\dim_{\mathrm{Min}}(M) := \limsup_{r \to 0} \dim_r(M).$$

Insights

If $\dim_{\mathrm{Min}}(M) = d$, we expect that

$$\log \mathrm{cv}_M(r) \sim_{r \to 0} -d \log r$$

Regression

We can regress

$$r \mapsto \log \mathrm{cv}_{\mathcal{X}}(r)$$

Minkowski / Capacity Dimension

Two-Scales Estimation

Instead of regression, **kegl2002intrinsic** uses the fact that for all small $r_1 < r_2$,

$$\frac{\log \mathrm{cv}_M(r_1) - \log \mathrm{cv}_M(r_2)}{\log r_2 - \log r_1} \simeq \frac{-d \log r_1 - (-d \log r_2)}{\log r_2 - \log r_1} = d.$$

Limitations

Still a choice of bandwidth(s) parameter(s).

Costly to compute directly on data (involves covering numbers).

Wayaround

Try to observe the dimension indirectly on a simpler object.

Partial Coverings

weed2019sharp introduce the Wasserstein dimension.

Idea

When working with measures instead of sets, it is convenient to be able to ignore a small fraction of the mass.

Definition

The (r, τ)-covering number of a probability measure P on $\mathbb{R}^p$ is

$$\mathrm{cv}_P(r, \tau) := \min \{ \mathrm{cv}_S(r) \,|\, P(S) \geq 1 - \tau \} .$$

Its (r, τ)-dimension is

$$\dim_{r,\tau}(P) := \frac{\log \mathrm{cv}_P(r, \tau)}{-\log r} .$$

Definition (weed2019sharp)

The upper and lower Wasserstein dimensions of P are respectively

$$\overline{d}^{(p)}(P) := \inf\left\{ s > 2p \,\middle|\, \limsup_{r\to 0}\ \dim_{r, r^{\frac{sp}{s-2p}}}(P) \le s \right\}$$

$$\underline{d}^{(p)}(P) := \lim_{\tau\to 0} \liminf_{r\to 0} \dim_{r,\tau}(P).$$

Definition (weed2019sharp)

The upper and lower Wasserstein dimensions of P are respectively

$$\overline{d}^{(p)}(P) := \inf\left\{ s > 2p \,\middle|\, \limsup_{r \to 0} \; \dim_{r, r^{\frac{sp}{s-2p}}}(P) \leq s \right\}$$

$$\underline{d}^{(p)}(P) := \lim_{\tau \to 0} \liminf_{r \to 0} \dim_{r,\tau}(P).$$

Links with other dimensions

If $P(B(x,r)) \asymp r^d$, then $\overline{d}^{(p)}(P) = d = \underline{d}^{(p)}(P)$

Generalizable to arbitrary ambient metric space.

Wasserstein Dimension

Convergence of the Empirical Distribution

Theorem (weed2019sharp)

Let $p \geq 1$. Assume that $x_1, \ldots, x_n \sim_{iid} P$ on $\mathbb{R}^d$. and write $P_n := n^{-1} \sum_{i=1}^{n} \delta_{x_i}$ for the empirical measure.

$$\text{If } s > \overline{\dim}^{(p)}(P), \text{ then } \mathbb{E}\left[\mathrm{W}_p(P, P_n)\right] \lesssim n^{-1/s}.$$

$$\text{If } t < \underline{\dim}^{(p)}(P), \text{ then } \mathbb{E}\left[\mathrm{W}_p(P, P_n)\right] \gtrsim n^{-1/t}.$$

Wasserstein Dimension

Convergence of the Empirical Distribution

Theorem (weed2019sharp)

Let $p \geq 1$. Assume that $x_1, \ldots, x_n \sim_{iid} P$ on $\mathbb{R}^d$. and write $P_n := n^{-1} \sum_{i=1}^n \delta_{x_i}$ for the empirical measure.

If $s > \overline{\dim}^{(p)}(P)$, then $\mathbb{E}\left[\mathrm{W}_p(P, P_n)\right] \lesssim n^{-1/s}$.

If $t < \underline{\dim}^{(p)}(P)$, then $\mathbb{E}\left[\mathrm{W}_p(P, P_n)\right] \gtrsim n^{-1/t}$.

The upper bound comes from a spatial dyadic decomposition.

The lower bound holds for all distribution P_n supported on n Dirac. It arises from a quantization argument.

Fine results on $W_p(P, P_n)$ in **dedecker2019behavior**.

block2021intrinsic leverage this sharp behavior as follows.

Bootstrap-Style Method

Given $0 < \alpha < 1$, subsample: (need $2(1+\alpha)n$ independent sample)

P_n, P'_n each arising from n observations each

$P_{\alpha n}, P'_{\alpha n}$ each arising from $\alpha n < n$ observations each

As

$$\mathrm{W}_1(P_{[\alpha]n}, P'_{[\alpha]n}) \asymp \mathrm{W}_1(P, P_{[\alpha]n}) \asymp \big(1/([\alpha]n)\big)^{1/d},$$

take

$$\hat{d}_{\mathrm{W}} := \frac{\log \alpha}{\log \mathrm{W}_1(P_n, P'_n) - \log \mathrm{W}_1(P_{\alpha n}, P'_{\alpha n})}$$

Which Wasserstein Metric?

Possibility to use the (estimated) geodesic metric in Wasserstein.

Geometric inference

Take a step back

Throughout, we have tried to embed points $\mathcal{X} \subset \mathbb{R}^p$ to $\mathcal{Y} \subset \mathbb{R}^d$ while preserving the geometry of $\mathcal{X}$.

If we assume that $\mathcal{X} \subset M$ are sample from a submanifold $M \subset \mathbb{R}^p$:

Preserving the geometry of $\mathcal{X}$

$$\Leftrightarrow$$

$$\mathrm{d}_M(x_i, x_j) \simeq \|y_i - y_j\|$$

Take a step back

Throughout, we have tried to embed points $\mathcal{X} \subset \mathbb{R}^p$ to $\mathcal{Y} \subset \mathbb{R}^d$ while preserving the geometry of $\mathcal{X}$.

If we assume that $\mathcal{X} \subset M$ are sample from a submanifold $M \subset \mathbb{R}^p$:

Preserving the geometry of $\mathcal{X}$

$$\Leftrightarrow$$

$$\mathrm{d}_M(x_i, x_j) \simeq \|y_i - y_j\|$$

The geodesic distance on M (or shortest-path distance)

$$\begin{aligned} \mathrm{d}_M :\ M \times M &\longrightarrow \mathbb{R}_{\geq 0} \cup \{\infty\} \\ (x, y) &\longmapsto \inf_{\substack{\gamma_{x \to y} \subset M \\ \mathcal{C}^1 \text{ curve}}} \int \|\gamma'_{x \to y}\| \end{aligned}$$

Take a step back

Throughout, we have tried to embed points $\mathcal{X} \subset \mathbb{R}^p$ to $\mathcal{Y} \subset \mathbb{R}^d$ while preserving the geometry of $\mathcal{X}$.

If we assume that $\mathcal{X} \subset M$ are sample from a submanifold $M \subset \mathbb{R}^p$:

Preserving the geometry of $\mathcal{X}$

$$\Leftrightarrow$$

$$\mathrm{d}_M(x_i, x_j) \simeq \|y_i - y_j\|$$

The geodesic distance on M (or shortest-path distance)

$$\begin{aligned} \mathrm{d}_M : \; M \times M &\longrightarrow \mathbb{R}_{\geq 0} \cup \{\infty\} \\ (x, y) &\longmapsto \inf_{\substack{\gamma_{x\to y} \subset M \\ \mathcal{C}^1 \text{ curve}}} \int \|\gamma'_{x\to y}\| \end{aligned}$$

What about only estimating d_M without embedding?

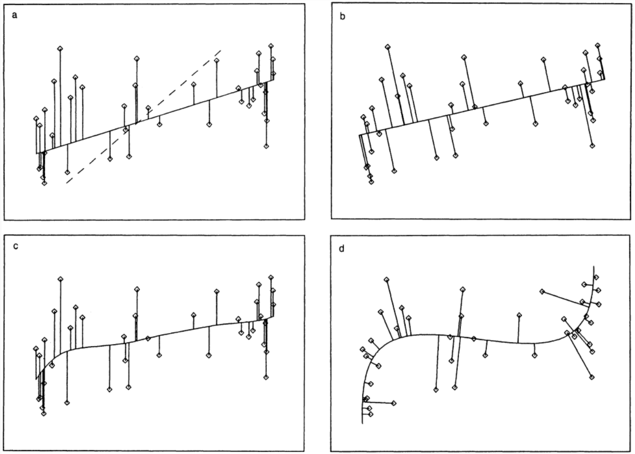


Figure 1. (a) The linear regression line minimizes the sum of squared deviations in the response variable. (b) The principal-component line minimizes the sum of squared deviations in all of the variables. (c) The smooth regression curve minimizes the sum of squared deviations in the response variable, subject to smoothness constraints. (d) The principal curve minimizes the sum of squared deviations in all of the variables, subject to smoothness constraints.

Figure 1: from **hastie1989principal**

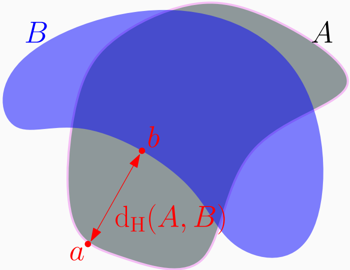
Hausdorff Distance

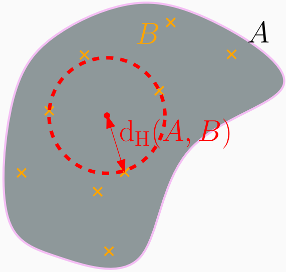
Definition (Hausdorff Distance)

The *Hausdorff distance* between two compact sets $A, B \subset \mathbb{R}^D$ is

$$\mathrm{d_H}(A, B) = \|\mathrm{d}(\cdot, A) - \mathrm{d}(\cdot, B)\|_\infty,$$

where $\mathrm{d}(x, C) := \inf_{c \in C} \|x - c\|$ is the distance to $C \subset \mathbb{R}^D$.

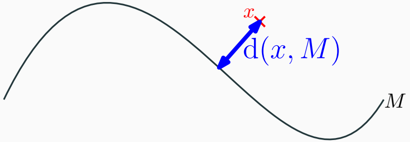




Disambiguation

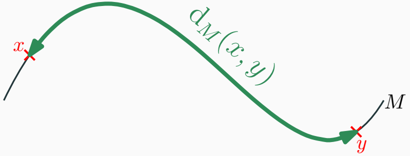
– The *distance function* to M : (used to identify sets as functions)

$$\begin{aligned} \mathrm{d}(\cdot, M)\colon\ \mathbb{R}^D &\longrightarrow \mathbb{R}_{\geq 0} \\ x &\longmapsto \min_{p \in M} \|x - p\| \end{aligned}$$



– The *geodesic distance* on M : (used to characterize the geometry of sets)

$$\begin{aligned} \mathrm{d}_M\colon\ M \times M &\longrightarrow \mathbb{R}_{\geq 0} \cup \{\infty\} \\ (x, y) &\longmapsto \inf_{\substack{\gamma_{x \to y} \subset M \\ \mathcal{C}^1 \text{ curve}}} \int \|\gamma'_{x \to y}\| \end{aligned}$$



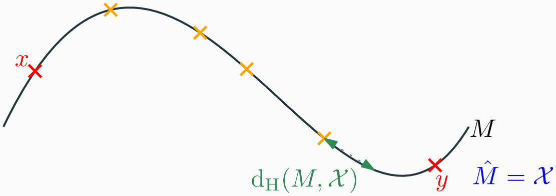
From Manifold Estimation to Metric Learning

Theorem (Aamari, Berenfeld, Levrard – 2023)

Assume that $M \subset \mathbb{R}^D$ is $\mathcal{C}^2$-smooth. Then there exists $\mathrm{rch}_M > 0$ such that for all $\hat{M} \subset \mathbb{R}^D$ such that $\mathrm{d_H}(M, \hat{M}) \leq \varepsilon < \mathrm{rch}_M/2$,

$$\sup_{x \neq y \in M} \left| \mathrm{d}_M(x,y) - \mathrm{d}_{(\hat{M})^\varepsilon}(x,y) \right| \lesssim \varepsilon,$$

where $(\hat{M})^\varepsilon := \{u \in \mathbb{R}^D \mid \mathrm{d}(u, \hat{M}) \leq \varepsilon\}$ is the ε-offset of M.



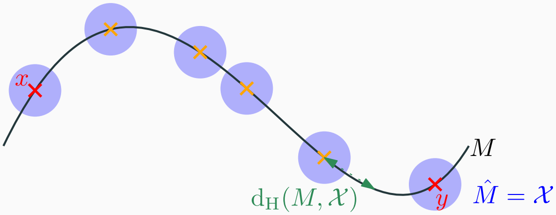
From Manifold Estimation to Metric Learning

Theorem (Aamari, Berenfeld, Levrard – 2023)

Assume that $M \subset \mathbb{R}^D$ *is* $\mathcal{C}^2$*-smooth. Then there exists* $\mathrm{rch}_M > 0$ *such that for all* $\hat{M} \subset \mathbb{R}^D$ *such that* $\mathrm{d_H}(M, \hat{M}) \leq \varepsilon < \mathrm{rch}_M/2$,

$$\sup_{x \neq y \in M} \left| \mathrm{d}_M(x,y) - \mathrm{d}_{(\hat{M})^\varepsilon}(x,y) \right| \lesssim \varepsilon,$$

where $(\hat{M})^\varepsilon := \{u \in \mathbb{R}^D \mid \mathrm{d}(u, \hat{M}) \leq \varepsilon\}$ *is the* ε*-offset of* M.



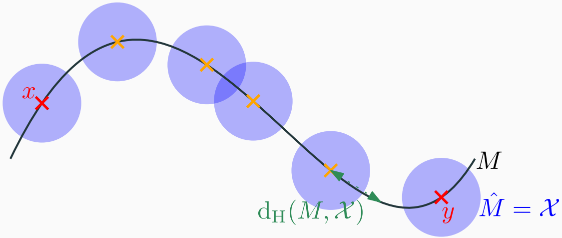
From Manifold Estimation to Metric Learning

Theorem (Aamari, Berenfeld, Levrard – 2023)

Assume that $M \subset \mathbb{R}^D$ *is* $\mathcal{C}^2$*-smooth. Then there exists* $\mathrm{rch}_M > 0$ *such that for all* $\hat{M} \subset \mathbb{R}^D$ *such that* $\mathrm{d_H}(M, \hat{M}) \leq \varepsilon < \mathrm{rch}_M/2$,

$$\sup_{x \neq y \in M} \left| \mathrm{d}_M(x,y) - \mathrm{d}_{(\hat{M})^\varepsilon}(x,y) \right| \lesssim \varepsilon,$$

where $(\hat{M})^\varepsilon := \{u \in \mathbb{R}^D \mid \mathrm{d}(u, \hat{M}) \leq \varepsilon\}$ *is the* ε*-offset of* M.



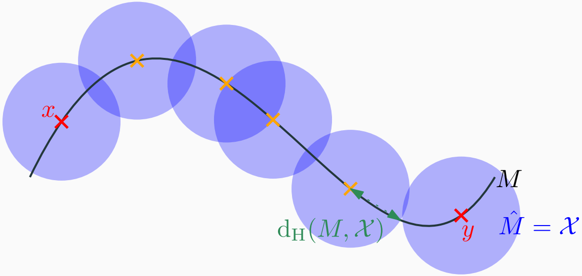
From Manifold Estimation to Metric Learning

Theorem (Aamari, Berenfeld, Levrard – 2023)

Assume that $M \subset \mathbb{R}^D$ *is* $\mathcal{C}^2$*-smooth. Then there exists* $\mathrm{rch}_M > 0$ *such that for all* $\hat{M} \subset \mathbb{R}^D$ *such that* $\mathrm{d_H}(M, \hat{M}) \leq \varepsilon < \mathrm{rch}_M/2$*,*

$$\sup_{x \neq y \in M} \left| \mathrm{d}_M(x, y) - \mathrm{d}_{(\hat{M})^\varepsilon}(x, y) \right| \lesssim \varepsilon,$$

where $(\hat{M})^\varepsilon := \{u \in \mathbb{R}^D \mid \mathrm{d}(u, \hat{M}) \leq \varepsilon\}$ *is the* ε*-offset of* M*.*



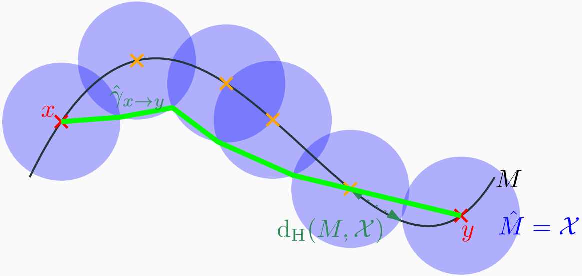
From Manifold Estimation to Metric Learning

Theorem (Aamari, Berenfeld, Levrard – 2023)

Assume that $M \subset \mathbb{R}^D$ *is* $\mathcal{C}^2$*-smooth. Then there exists* $\mathrm{rch}_M > 0$ *such that for all* $\hat{M} \subset \mathbb{R}^D$ *such that* $\mathrm{d_H}(M, \hat{M}) \leq \varepsilon < \mathrm{rch}_M/2$*,*

$$\sup_{x \neq y \in M} \left| \mathrm{d}_M(x,y) - \mathrm{d}_{(\hat{M})^\varepsilon}(x,y) \right| \lesssim \varepsilon,$$

where $(\hat{M})^\varepsilon := \{u \in \mathbb{R}^D \mid \mathrm{d}(u, \hat{M}) \leq \varepsilon\}$ *is the* ε*-offset of* M*.*



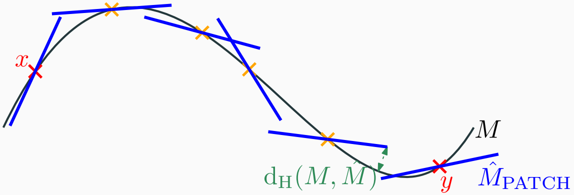
From Manifold Estimation to Metric Learning

Theorem (Aamari, Berenfeld, Levrard – 2023)

Assume that $M \subset \mathbb{R}^D$ *is* $\mathcal{C}^2$*-smooth. Then there exists* $\mathrm{rch}_M > 0$ *such that for all* $\hat{M} \subset \mathbb{R}^D$ *such that* $\mathrm{d_H}(M, \hat{M}) \leq \varepsilon < \mathrm{rch}_M/2$,

$$\sup_{x \neq y \in M} \left| \mathrm{d}_M(x, y) - \mathrm{d}_{(\hat{M})^\varepsilon}(x, y) \right| \lesssim \varepsilon,$$

where $(\hat{M})^\varepsilon := \{u \in \mathbb{R}^D \mid \mathrm{d}(u, \hat{M}) \leq \varepsilon\}$ *is the* ε*-offset of* M.



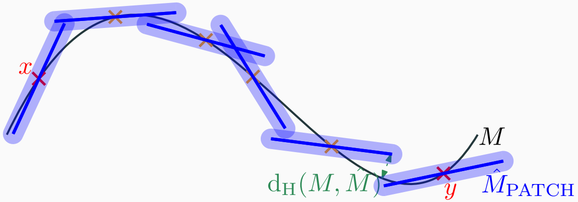
From Manifold Estimation to Metric Learning

Theorem (Aamari, Berenfeld, Levrard – 2023)

Assume that $M \subset \mathbb{R}^D$ *is* $\mathcal{C}^2$*-smooth. Then there exists* $\mathrm{rch}_M > 0$ *such that for all* $\hat{M} \subset \mathbb{R}^D$ *such that* $\mathrm{d_H}(M, \hat{M}) \leq \varepsilon < \mathrm{rch}_M/2$,

$$\sup_{x \neq y \in M} \left| \mathrm{d}_M(x,y) - \mathrm{d}_{(\hat{M})^\varepsilon}(x,y) \right| \lesssim \varepsilon,$$

where $(\hat{M})^\varepsilon := \{u \in \mathbb{R}^D \mid \mathrm{d}(u, \hat{M}) \leq \varepsilon\}$ *is the* ε*-offset of* M.



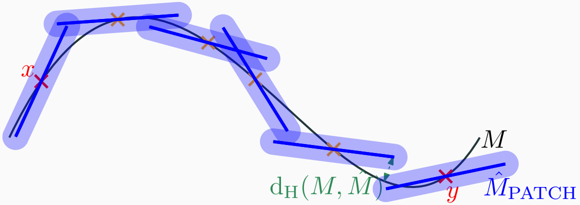
From Manifold Estimation to Metric Learning

Theorem (Aamari, Berenfeld, Levrard – 2023)

Assume that $M \subset \mathbb{R}^D$ *is* $\mathcal{C}^2$*-smooth. Then there exists* $\mathrm{rch}_M > 0$ *such that for all* $\hat{M} \subset \mathbb{R}^D$ *such that* $\mathrm{d_H}(M, \hat{M}) \leq \varepsilon < \mathrm{rch}_M/2$,

$$\sup_{x \neq y \in M} \left| \mathrm{d}_M(x, y) - \mathrm{d}_{(\hat{M})^\varepsilon}(x, y) \right| \lesssim \varepsilon,$$

where $(\hat{M})^\varepsilon := \{u \in \mathbb{R}^D \mid \mathrm{d}(u, \hat{M}) \leq \varepsilon\}$ *is the* ε*-offset of* M.



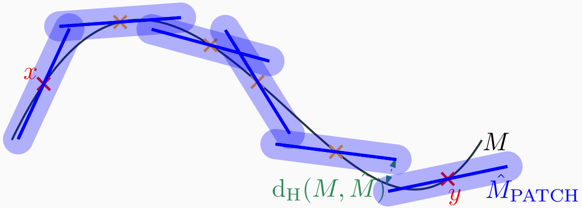
From Manifold Estimation to Metric Learning

Theorem (Aamari, Berenfeld, Levrard – 2023)

Assume that $M \subset \mathbb{R}^D$ *is* $\mathcal{C}^2$*-smooth. Then there exists* $\mathrm{rch}_M > 0$ *such that for all* $\hat{M} \subset \mathbb{R}^D$ *such that* $\mathrm{d_H}(M, \hat{M}) \leq \varepsilon < \mathrm{rch}_M/2$,

$$\sup_{x \neq y \in M} \left| \mathrm{d}_M(x, y) - \mathrm{d}_{(\hat{M})^\varepsilon}(x, y) \right| \lesssim \varepsilon,$$

where $(\hat{M})^\varepsilon := \{u \in \mathbb{R}^D \mid \mathrm{d}(u, \hat{M}) \leq \varepsilon\}$ *is the* ε*-offset of* M.



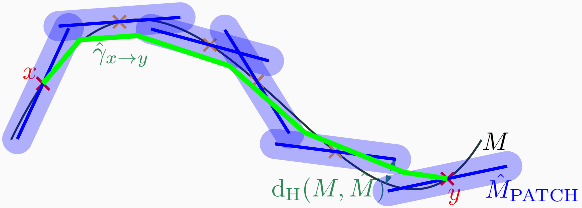
From Manifold Estimation to Metric Learning

Theorem (Aamari, Berenfeld, Levrard – 2023)

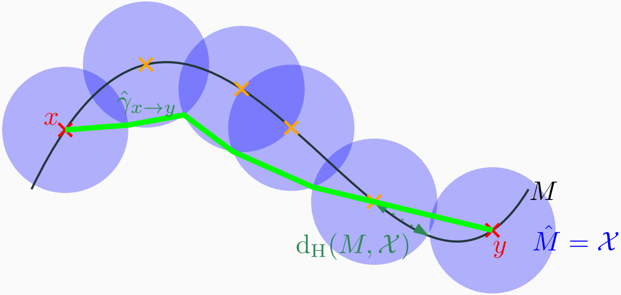
Assume that $M \subset \mathbb{R}^D$ *is* $\mathcal{C}^2$*-smooth. Then there exists* $\mathrm{rch}_M > 0$ *such that for all* $\hat{M} \subset \mathbb{R}^D$ *such that* $\mathrm{d_H}(M, \hat{M}) \leq \varepsilon < \mathrm{rch}_M/2$*,*

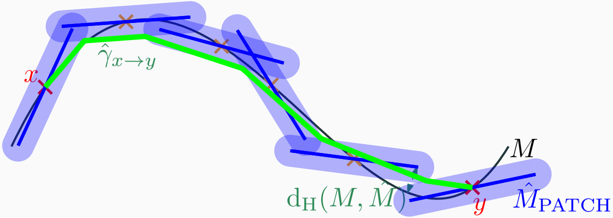
$$\sup_{x \neq y \in M} \left| \mathrm{d}_M(x,y) - \mathrm{d}_{(\hat{M})^\varepsilon}(x,y) \right| \lesssim \varepsilon,$$

where $(\hat{M})^\varepsilon := \{u \in \mathbb{R}^D \mid \mathrm{d}(u, \hat{M}) \leq \varepsilon\}$ *is the* ε*-offset of* M*.*

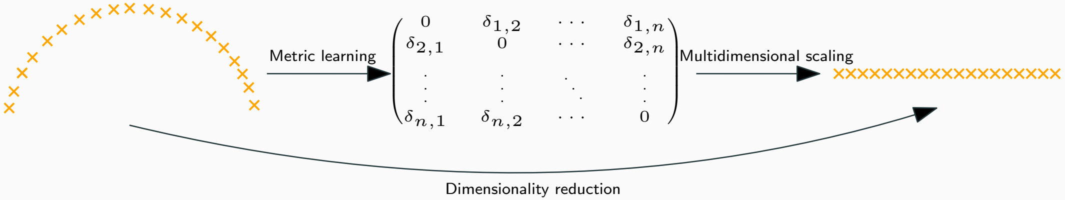


Better Manifold Estimation, Better Metric Learning

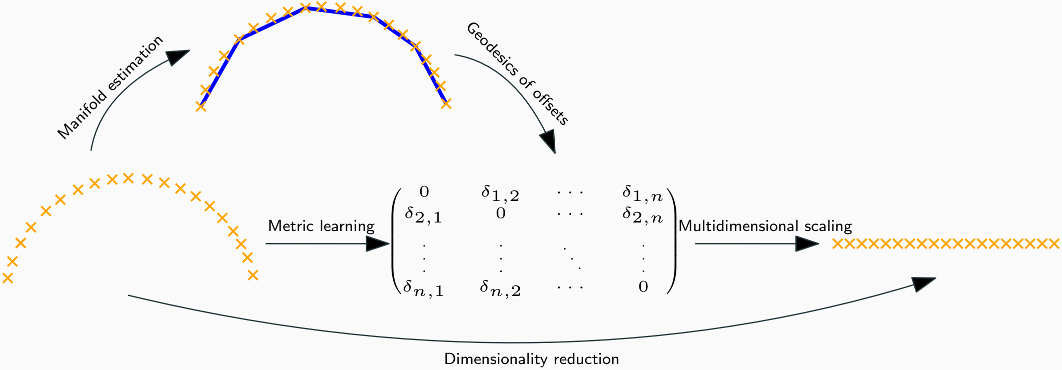




Manifold estimation ⇒ Dimensionality reduction

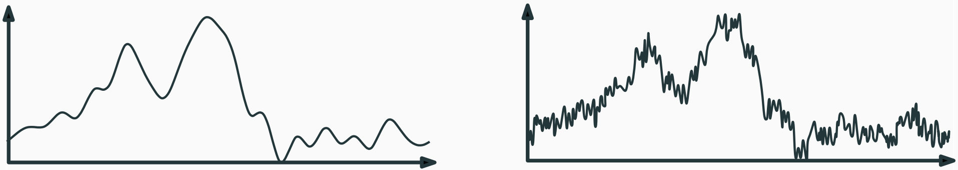


Manifold estimation ⇒ Dimensionality reduction



Regularity in nonparametric geometric problems

Regularity

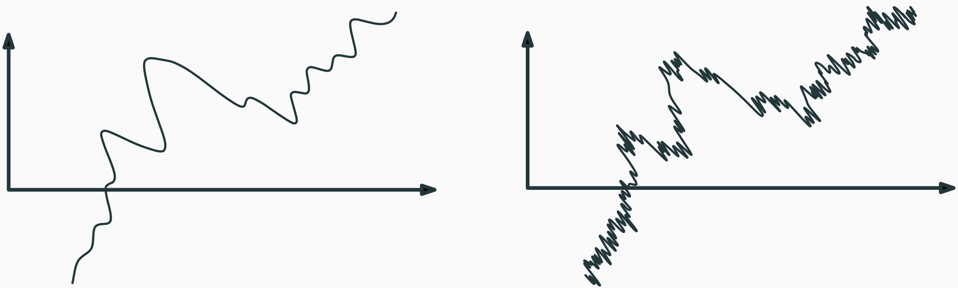


Usual regularity classes (Hölder, Sobolev, Besov) control increments

$$\|f(x) - f(y)\| \leq L \|x - y\|^{\beta} .$$

(L, β) drives the difficulty of the statistical problem.

Regularity Without Coordinates?



Usual regularity classes (Hölder, Sobolev, Besov) control increments

$$\|f(x) - f(y)\| \leq L \|x - y\|^{\beta} .$$

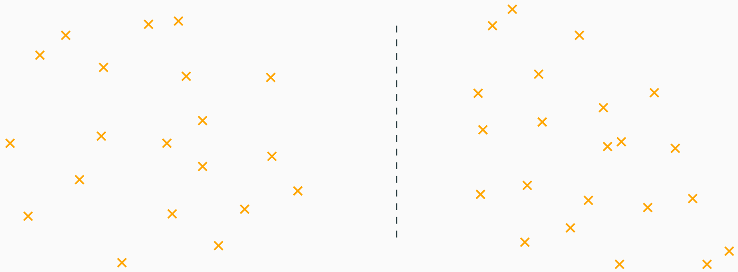
(L, β) drives the difficulty of the statistical problem.

Without natural coordinates, "$\|f(x) - f(y)\|$" $= ?$

Support Estimation

Data: A n-sample $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$.

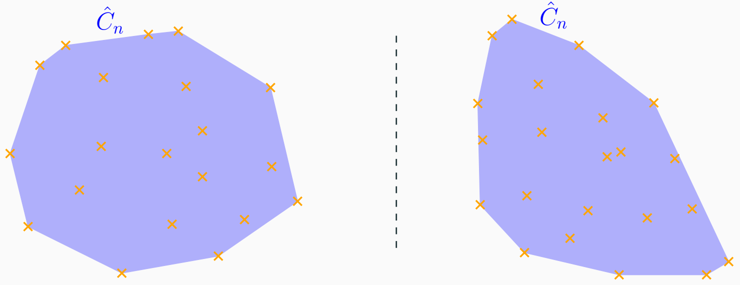
Goal: Estimate the set $C = \mathrm{Support}(P) = \bigcap_{\substack{K \subset \mathbb{R}^D \text{ closed} \\ P(K)=1}} K$.



Support Estimation

Data: A n-sample $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$.

Goal: Estimate the set $C = \mathrm{Support}(P) = \bigcap_{\substack{K \subset \mathbb{R}^D \text{ closed} \\ P(K)=1}} K$.



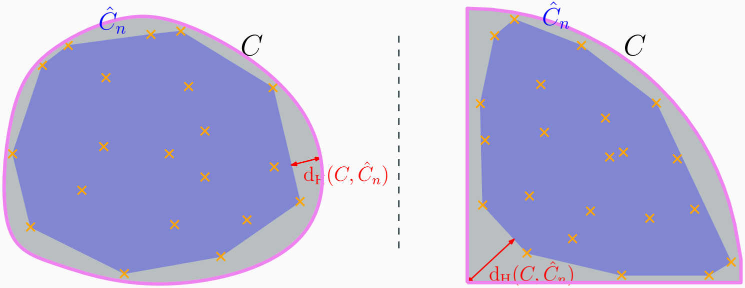
If we know (by advance) that C is convex, a good candidate is

$$\hat{C}_n := \mathrm{Conv}(\{X_1, \ldots, X_n\}).$$

Support Estimation

Data: A n-sample $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$.

Goal: Estimate the set $C = \mathrm{Support}(P) = \bigcap_{\substack{K \subset \mathbb{R}^D \text{ closed} \\ P(K)=1}} K$.



If we know (by advance) that C is convex, a good candidate is

$$\hat{C}_n := \mathrm{Conv}(\{X_1, \ldots, X_n\}).$$

Support Estimation: Convex Case(s)

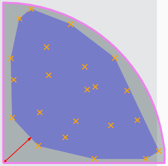
Theorem (Dümbgen, Walther – 1996)

Assume that $P = \mathrm{Unif}_C$ *is uniform over the convex set* $C \subset \mathbb{R}^D$. *Write*

$$\mathbb{X}_n := \{X_1, \ldots, X_n\}, \quad \text{and} \quad \hat{C}_n = \mathrm{Conv}(\mathbb{X}_n).$$

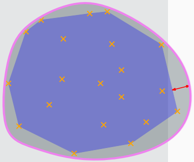
– *Then,*

$$\mathrm{d_H}(C, \mathbb{X}_n) \asymp \mathrm{d_H}(C, \hat{C}_n) = O\left(\frac{\log n}{n}\right)^{\frac{1}{D}} \quad a.s.$$

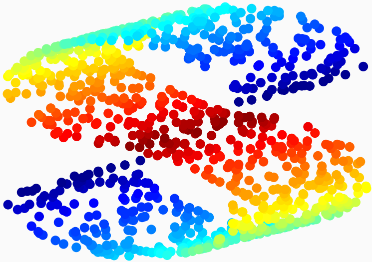


– *If in addition* ∂C *is* $\mathcal{C}^2$,

$$\mathrm{d_H}(C, \hat{C}_n) = O\left(\frac{\log n}{n}\right)^{\frac{2}{D+1}} \quad a.s.$$



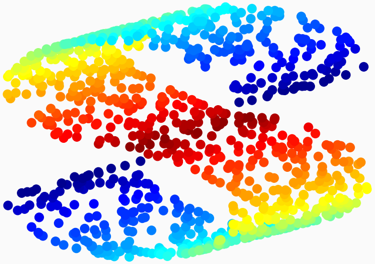
Beyond Convexity



How to model the support of these data?

- Low-dimensional and curved $\rightarrow$ Submanifold of $\mathbb{R}^D$.
- Not convex, but *locally* around it the projection uniquely defined.

Beyond Convexity



How to model the support of these data?

- Low-dimensional and curved $\rightarrow$ Submanifold of $\mathbb{R}^D$.
- Not convex, but *locally* around it the projection uniquely defined.

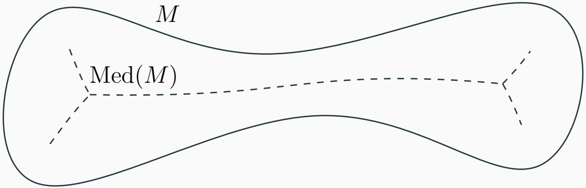
Reminder: For a closed set $C \subset \mathbb{R}^D$,

$C \subset \mathbb{R}^D$ is convex $\Leftrightarrow$ Every $z \in \mathbb{R}^D$ has a unique nearest neighbor on C
i.e. $\exists!\ \pi_C(z) \in C$ with $\|z - \pi_C(z)\| = \mathrm{d}(z, C)$.

Medial Axis

The **medial axis** of $M \subset \mathbb{R}^D$ is the set of points that have ≥ 2 nearest neighbors on M:

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



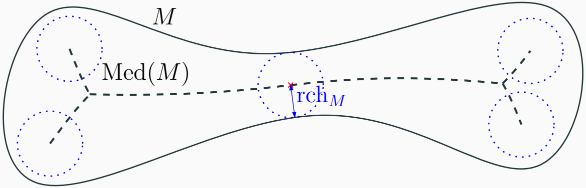
Medial axis of a curve

Reach

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

$$\mathrm{rch}_M := \inf_{x \in M} \mathrm{d}\left(x, \mathrm{Med}(M)\right),$$

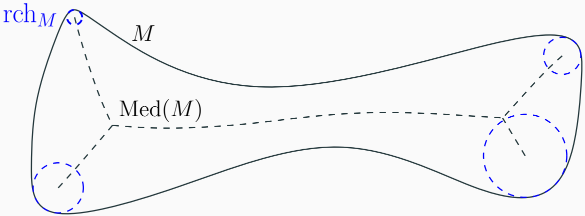
where for all $x \in \mathbb{R}^D$, $\mathrm{d}(x, K) := \inf_{p \in K} \|x - p\|$.



One can also flip the formula:

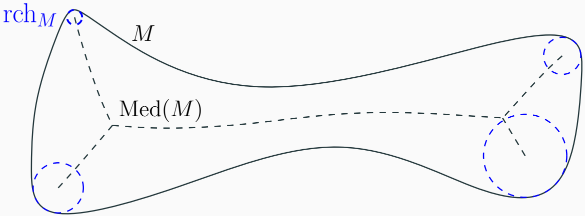
$$\mathrm{rch}_M = \inf_{z \in \mathrm{Med}(M)} \mathrm{d}(z, M).$$

Local Regularity



High curvature $\Leftrightarrow$ Small radius of curvature $\Rightarrow$ $\mathrm{rch}_M \ll 1$.

Local Regularity



High curvature $\Leftrightarrow$ Small radius of curvature $\Rightarrow$ $\mathrm{rch}_M \ll 1$.

Proposition (Federer – 1959, Niyogi et al. – 2006)

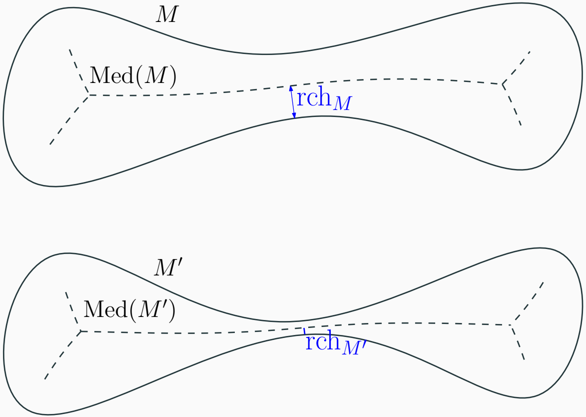
Let II^M_x denote the second fundamental form of M. For all unit tangent vector $v \in T_x M$,

$$\left\| II^M_x(v, v) \right\| \leq 1/\mathrm{rch}_M .$$

As a consequence, the sectional curvatures κ of M satisfy

$$-2/\mathrm{rch}^2_M \leq \kappa \leq 1/\mathrm{rch}^2_M .$$

Global Regularity



Narrow bottleneck structure $\Rightarrow$ $\mathrm{rch}_M \ll 1.$

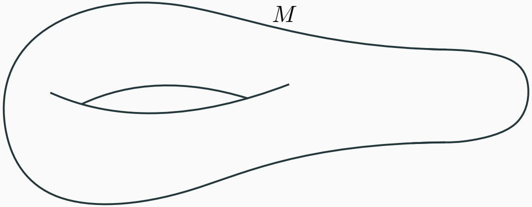
Noiseless manifold estimation

Boundariless Statistical Model

$X_1, \dots, X_n \overset{i.i.d.}{\sim} P$, where $M = \mathrm{Support}(P) \subset \mathbb{R}^D$ satisfies:

- M is a compact connected d-dimensional submanifold,
- M has **no boundary**,
- $\mathrm{rch}_M \geq \mathrm{rch}_{min} > 0$,
- P is (almost) the uniform distribution on M.

The set of distributions satisfying these conditions is denoted by $\mathcal{P}$.



A Reconstruction Theorem

Theorem (Aamari, Levrard – 2018)

If $P \in \mathcal{P}$, one can compute an estimator $\hat{M}_{\texttt{PATCH}}$ based on data points $\mathbb{X}_n$ such that w.h.p.,

$$\mathrm{d_H}\left(M, \hat{M}_{\texttt{PATCH}}\right) \leq C \left(\frac{\log n}{n}\right)^{2/d}.$$

Here, $C = C_{\mathrm{rch}_{min},d}$ does not depend on the ambient dimension D.

A Reconstruction Theorem

Theorem (Aamari, Levrard – 2018)

If $P \in \mathcal{P}$, one can compute an estimator $\hat{M}_{\mathtt{PATCH}}$ based on data points $\mathbb{X}_n$ such that w.h.p.,

$$\mathrm{d_H}\left(M, \hat{M}_{\mathtt{PATCH}}\right) \leq C \left(\frac{\log n}{n}\right)^{2/d}.$$

Here, $C = C_{\mathrm{rch}_{min},d}$ does not depend on the ambient dimension D.

$\rightarrow$ Other estimators achieving the same Hausdorff rate:

- Empirical risk manifold minimizer [Genovese *et al.* – 2012]
- Local Tangent Delaunay triangulation [Aamari & Levrard – 2019]
- Local convex hulls [Divol – 2020]

Ingredient I: Approximation Theory

Theorem (Aamari, Levrard – 2019)

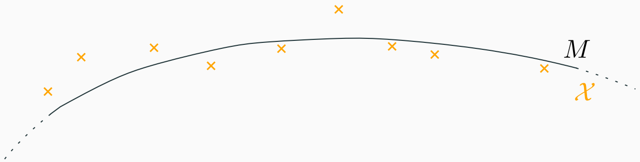
For $\Delta \lesssim \mathrm{rch}_{\min}$, assume that we have a point cloud $\mathcal{X} \subset \mathbb{R}^D$ that is:

- *close to M:* $\max_{x \in \mathcal{X}} \mathrm{d}(x, M) \lesssim \Delta^2/\mathrm{rch}_{\min}$,
- *a covering of M:* $\sup_{p \in M} \mathrm{d}(p, \mathcal{X}) \lesssim \Delta$,

together with a family $\mathbb{T}_{\mathcal{X}}$ of linear spaces that

- *approximate tangent spaces:* $\max_{x \in \mathcal{X}} \angle\left(T_{\pi_M(x)} M, T_x\right) \lesssim \Delta/\mathrm{rch}_{\min}$

One can build a local linear estimator $\hat{M}_{\texttt{PATCH}}(\mathcal{X}, \mathbb{T}_{\mathcal{X}})$ such that $\mathrm{d_H}\left(M, \hat{M}_{\texttt{PATCH}}\right) \lesssim \Delta^2/\mathrm{rch}_{\min}$.



Ingredient I: Approximation Theory

Theorem (Aamari, Levrard – 2019)

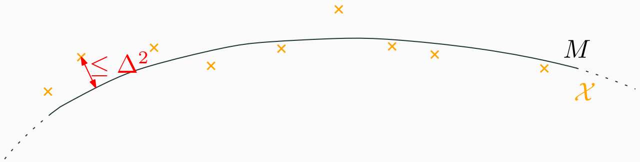
For $\Delta \lesssim \mathrm{rch}_{\min}$, assume that we have a point cloud $\mathcal{X} \subset \mathbb{R}^D$ that is:

- *close to M:* $\max_{x \in \mathcal{X}} \mathrm{d}(x, M) \lesssim \Delta^2/\mathrm{rch}_{\min}$,
- *a covering of M:* $\sup_{p \in M} \mathrm{d}(p, \mathcal{X}) \lesssim \Delta$,

together with a family $\mathbb{T}_{\mathcal{X}}$ of linear spaces that

- *approximate tangent spaces:* $\max_{x \in \mathcal{X}} \angle\left(T_{\pi_M(x)} M, T_x\right) \lesssim \Delta/\mathrm{rch}_{\min}$

One can build a local linear estimator $\hat{M}_{\mathtt{PATCH}}(\mathcal{X}, \mathbb{T}_{\mathcal{X}})$ such that $\mathrm{d_H}\left(M, \hat{M}_{\mathtt{PATCH}}\right) \lesssim \Delta^2/\mathrm{rch}_{\min}$.



Ingredient I: Approximation Theory

Theorem (Aamari, Levrard – 2019)

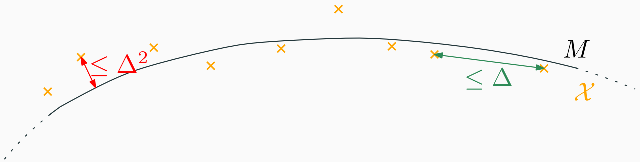
For $\Delta \lesssim \mathrm{rch}_{\min}$, assume that we have a point cloud $\mathcal{X} \subset \mathbb{R}^D$ that is:

- *close to M:* $\max_{x \in \mathcal{X}} \mathrm{d}(x, M) \lesssim \Delta^2 / \mathrm{rch}_{\min}$,
- *a covering of M:* $\sup_{p \in M} \mathrm{d}(p, \mathcal{X}) \lesssim \Delta$,

together with a family $\mathbb{T}_{\mathcal{X}}$ of linear spaces that

- *approximate tangent spaces:* $\max_{x \in \mathcal{X}} \angle\left(T_{\pi_M(x)} M, T_x\right) \lesssim \Delta / \mathrm{rch}_{\min}$

One can build a local linear estimator $\hat{M}_{\texttt{PATCH}}(\mathcal{X}, \mathbb{T}_{\mathcal{X}})$ such that $\mathrm{d_H}\left(M, \hat{M}_{\texttt{PATCH}}\right) \lesssim \Delta^2 / \mathrm{rch}_{\min}$.



Ingredient I: Approximation Theory

Theorem (Aamari, Levrard – 2019)

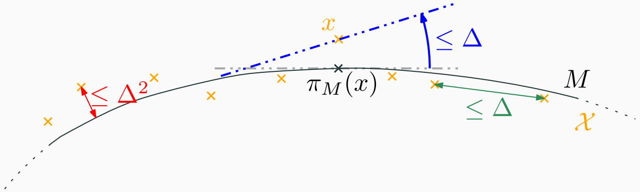
For $\Delta \lesssim \mathrm{rch}_{\min}$, assume that we have a point cloud $\mathcal{X} \subset \mathbb{R}^D$ that is:

- *close to M:* $\max_{x \in \mathcal{X}} \mathrm{d}(x, M) \lesssim \Delta^2 / \mathrm{rch}_{\min}$,
- *a covering of M:* $\sup_{p \in M} \mathrm{d}(p, \mathcal{X}) \lesssim \Delta$,

together with a family $\mathbb{T}_{\mathcal{X}}$ of linear spaces that

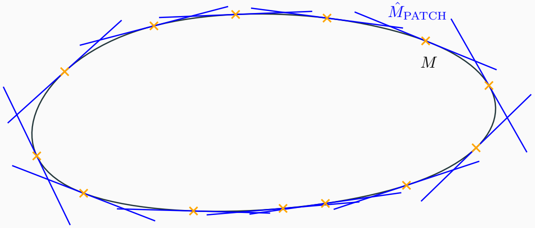
- *approximate tangent spaces:* $\max_{x \in \mathcal{X}} \angle\left(T_{\pi_M(x)} M, T_x\right) \lesssim \Delta / \mathrm{rch}_{\min}$

One can build a local linear estimator $\hat{M}_{\mathtt{PATCH}}(\mathcal{X}, \mathbb{T}_{\mathcal{X}})$ such that $\mathrm{d_H}\left(M, \hat{M}_{\mathtt{PATCH}}\right) \lesssim \Delta^2 / \mathrm{rch}_{\min}$.



Ingredient I: Approximation Theory

$$\hat{M}_{\mathrm{PATCH}} := \bigcup_{x \in \mathcal{X}} \mathrm{B}_{T_x}(0, \Delta).$$



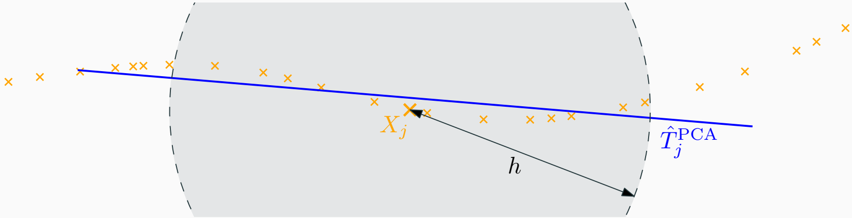
Ingredient II: Local PCA

Define $\hat{T}_j^{\mathrm{PCA}}$ to be a minimizer of

$$\hat{T}_j^{\mathrm{PCA}} \in \arg\min_{T} P_n^{(j)} \left[\|x - \pi_T(x)\|^2 \mathbf{1}_{\mathrm{B}(0,h)}(x) \right],$$

where:

- $P_n^{(j)}$ denotes the integration with respect to $\frac{1}{n} \sum_{\ell \neq j} \delta_{X_\ell - X_j}$,
- T ranges in the set of d-planes of $\mathbb{R}^D$.



Ingredient II: Local PCA

Define $\hat{T}_j^{\mathrm{PCA}}$ to be a minimizer of

$$\hat{T}_j^{\mathrm{PCA}} \in \arg\min_T P_n^{(j)} \left[\|x - \pi_T(x)\|^2 \mathbf{1}_{\mathrm{B}(0,h)}(x) \right],$$

where:

- $P_n^{(j)}$ denotes the integration with respect to $\frac{1}{n}\sum_{\ell \neq j} \delta_{X_\ell - X_j}$,
- T ranges in the set of d-planes of $\mathbb{R}^D$.

Theorem (Aamari, Levrard – 2019)

Picking $h \asymp (\log n/n)^{1/d}$, *then with high probability,*

$$\max_{1 \leq j \leq n} \angle(T_{X_j} M, \hat{T}_j^{\mathrm{PCA}}) \lesssim \left(\frac{\log n}{n} \right)^{1/d},$$

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

Manifold Estimation from Random Sample

Proposition (Aamari, Levrard – 2019)

An i.i.d. n-sample $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ of $P \in \mathcal{P}^{d,D}_{\mathrm{rch}_{\min}}$ fulfills:

- $\max_{X_j \in \mathbb{X}_n} \mathrm{d}(X_j, M) = 0$
- $\sup_{p \in M} \mathrm{d}(p, \mathbb{X}_n) \lesssim (\log n / n)^{1/d}$.

The family of d-planes $\hat{T}^{\mathrm{PCA}}_{\mathbb{X}_n}$ built from local PCA fulfills

$$\max_{X_j \in \mathbb{X}_n} \angle\left(T_{X_j} M, \hat{T}_{X_j}\right) \lesssim (\log n / n)^{1/d}.$$

$\Rightarrow$ *With high probability, we get precision:*

$$\varepsilon = \mathrm{d_H}\left(M, \hat{M}_{\mathrm{PATCH}}\right) \lesssim \left(\frac{\log n}{n}\right)^{2/d}.$$

This rate is minimax optimal

Optimality: Studying the Minimax Risk

The **minimax risk** over the statistical model $\mathcal{P}$ is

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

where $\hat{M}_n = \hat{M}_n(\mathbb{X}_n)$ ranges over all the estimators based on data $\mathbb{X}_n = \{X_1, \ldots, X_n\}$.

Optimality: Studying the Minimax Risk

The **minimax risk** over the statistical model $\mathcal{P}$ is

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

where $\hat{M}_n = \hat{M}_n(\mathbb{X}_n)$ ranges over all the estimators based on data $\mathbb{X}_n = \{X_1, \ldots, X_n\}$.

Proposition (Genovese et al – 2012)

For n large enough,

$$\inf_{\hat{M}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}\left[\mathrm{d_H}\left(M, \hat{M}_n\right)\right] \leq C \left(\frac{\log n}{n}\right)^{\frac{2}{d}},$$

where $C = C_{d,\mathrm{rch}_{\min}}$

Optimality: Studying the Minimax Risk

The **minimax risk** over the statistical model $\mathcal{P}$ is

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

where $\hat{M}_n = \hat{M}_n(\mathbb{X}_n)$ ranges over all the estimators based on data $\mathbb{X}_n = \{X_1, \ldots, X_n\}$.

Proposition (Genovese et al – 2012, Kim & Zhou – 2015)

For n large enough, (+ mild technical assumptions)

$$c\left(\frac{\log n}{n}\right)^{\frac{2}{d}} \leq \inf_{\hat{M}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}\left[\mathrm{d_H}\left(M, \hat{M}_n\right)\right] \leq C\left(\frac{\log n}{n}\right)^{\frac{2}{d}},$$

where $C = C_{d,\mathrm{rch}_{\min}}$ and $c = c_{\mathrm{rch}_{min}}$.

Lower Bound Technique: Le Cam's Lemma

Theorem (L. Le Cam)

For all $P_0, P_1 \in \mathcal{P}$,

$$\inf_{\hat{M}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}\left[\mathrm{d_H}\left(M, \hat{M}_n\right)\right] \geq \frac{1}{2}\mathrm{d_H}(M_0, M_1)\left(1 - \mathrm{TV}(P_0, P_1)\right)^n,$$

where

$$\mathrm{TV}(P_0, P_1) = \sup_{B \in \mathcal{B}(\mathbb{R}^D)} |P_0(B) - P_1(B)|$$

denotes the total variation distance between P_0 *and* P_1.

Lower Bound Technique: Le Cam's Lemma

Theorem (L. Le Cam)

For all $P_0, P_1 \in \mathcal{P}$,

$$\inf_{\hat{M}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}\left[\mathrm{d_H}\left(M, \hat{M}_n\right)\right] \geq \frac{1}{2}\mathrm{d_H}(M_0, M_1)\left(1 - \mathrm{TV}(P_0, P_1)\right)^n,$$

where

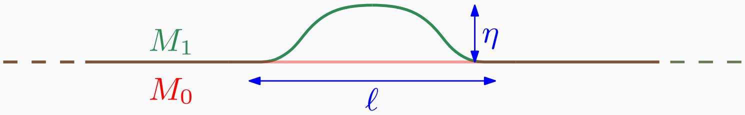
$$\mathrm{TV}(P_0, P_1) = \sup_{B \in \mathcal{B}(\mathbb{R}^D)} |P_0(B) - P_1(B)|$$

denotes the total variation distance between P_0 *and* P_1.

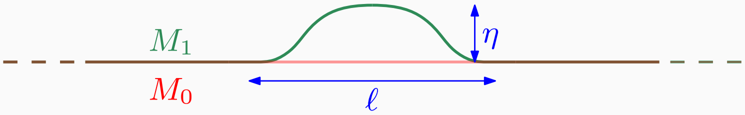
Deriving a good lower bound amounts to find P_0, P_1 such that:

- $P_0, P_1 \in \mathcal{P}$,
- $\mathrm{d_H}(M_0, M_1)$ is large,
- $\mathrm{TV}(P_0, P_1)$ is small.

Le Cam's Lemma Heuristic

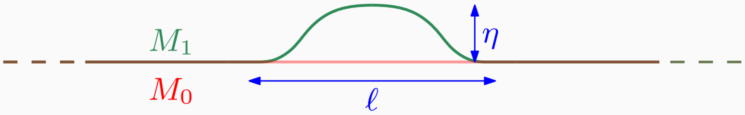


Le Cam's Lemma Heuristic



- P_0 and P_1 both belong to $\mathcal{P}$ as soon as $\eta \lesssim \ell^2$,
- $\mathrm{d_H}(M_0, M_1) \geq \eta$,
- $\mathrm{TV}(P_0, P_1) \lesssim \ell^d$.

Le Cam's Lemma Heuristic



- P_0 and P_1 both belong to $\mathcal{P}$ as soon as $\eta \lesssim \ell^2$,
- $\mathrm{d_H}(M_0, M_1) \geq \eta$,
- $\mathrm{TV}(P_0, P_1) \lesssim \ell^d$.

Hence, for $\eta \approx \ell^2$ and $\ell \approx (1/n)^{1/d}$,

$$\inf_{\hat{M}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}\left[\mathrm{d_H}\left(M, \hat{M}_n\right)\right] \gtrsim \eta \left(1 - \ell^d\right)^n \approx \ell^2 \left(1 - \ell^d\right)^n \approx (1/n)^{2/d}.$$

What if the Curve isn't Closed?

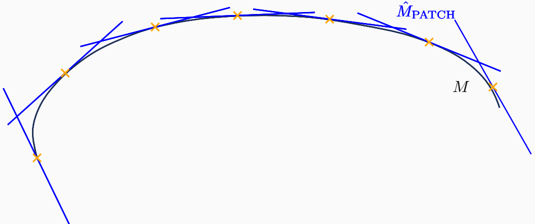
Perform local PCA at each point $X_j \in \mathbb{X}_n$:

$$\hat{T}_j \in \operatorname*{arg\,min}_{T \in \mathbb{G}^{D,d}} P_n^{(j)} \left[\|x - \pi_T(x)\|^2 \mathbf{1}_{\mathrm{B}(0,h)}(x) \right],$$

and take

$$\hat{M}_{\text{PATCH}} := \bigcup_{j=1}^{n} \mathrm{B}_{\hat{T}_j}(0, h).$$

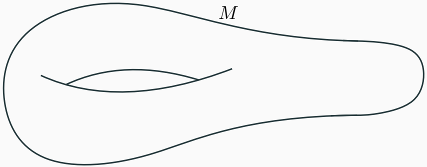
+ Local PCA still estimates tangent spaces up to angle $\lesssim (\log n/n)^{1/d}$.
- Nearby "boundary points", the patches extend too far away from M.

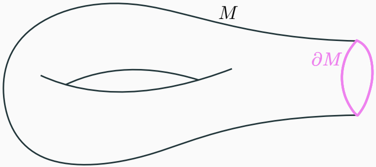


Boundary Manifold Model

We let $\mathcal{P}^\partial := \mathcal{P}^{d,D}_{\mathrm{rch}_{\min},\mathrm{rch}_{\partial,\min}}$ denote the set of distributions P over $\mathbb{R}^D$ such that

- Its support $M = \mathrm{supp}(P) \subset \mathbb{R}^D$ satisfies:
  - M is a $\mathcal{C}^2$ submanifold **with boundary**;
  - M has reach bounded away from zero $\mathrm{rch}_M \geq \mathrm{rch}_{\min} > 0$;
  - ∂M has reach bounded away from zero $\mathrm{rch}_{\partial M} \geq \mathrm{rch}_{\partial,\min} > 0$.

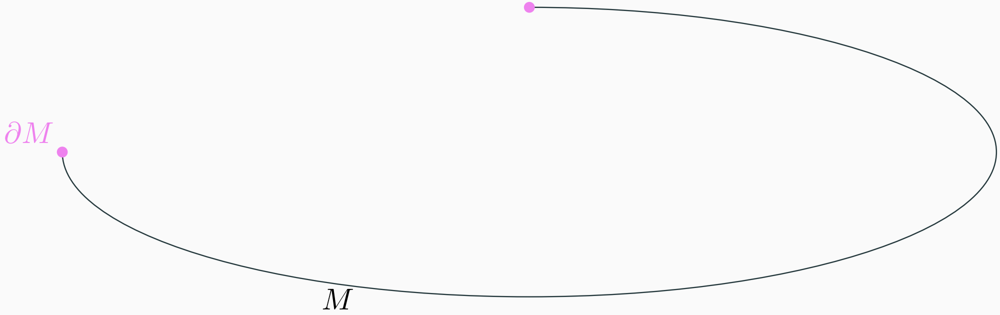




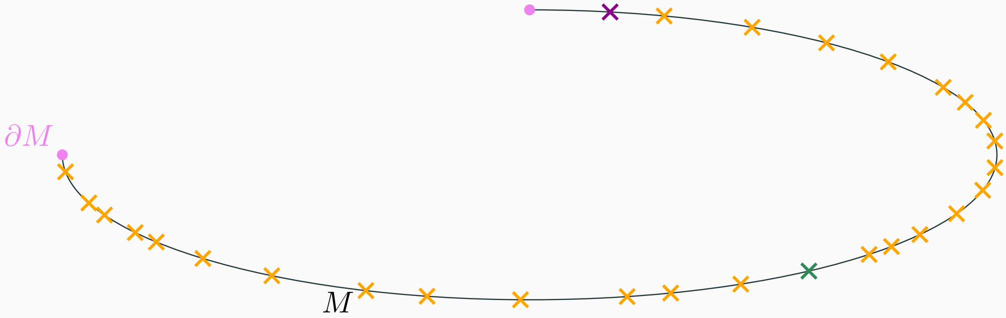
- P is roughly uniform on M:

$$f = \mathrm{d}P/\mathrm{dvol}_M \text{ exists and } f_{\min} \leq f \leq f_{\max}.$$

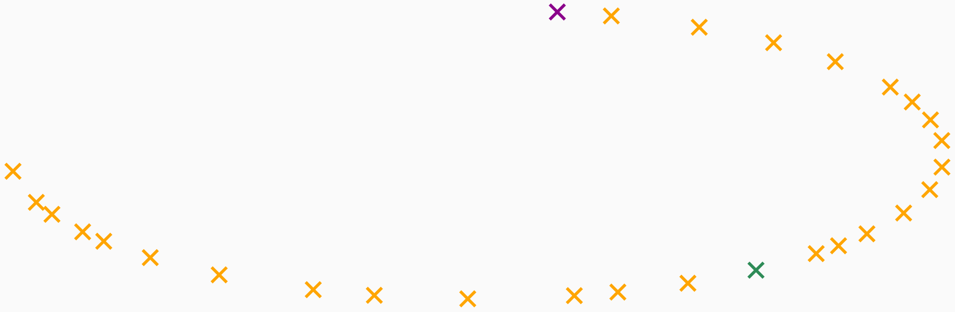
Insights on Boundary Point Detection



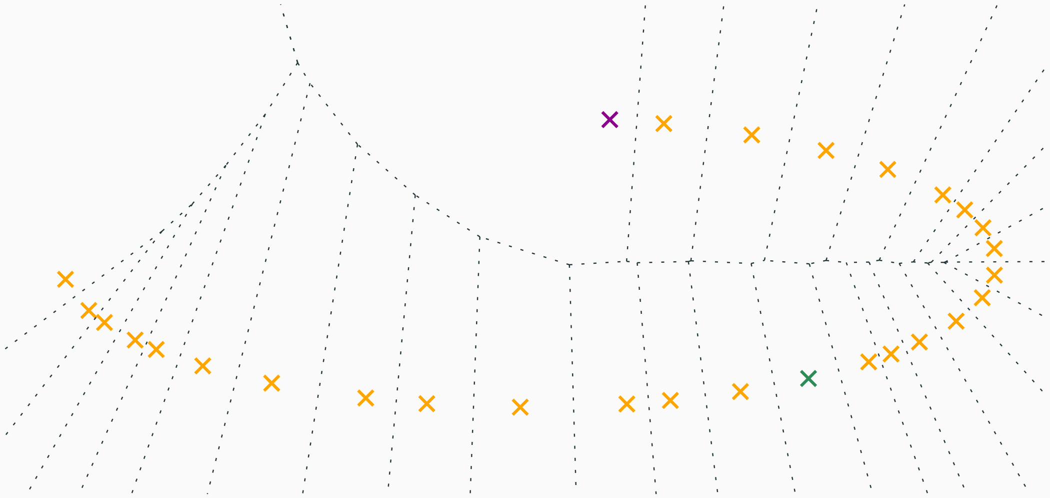
Insights on Boundary Point Detection



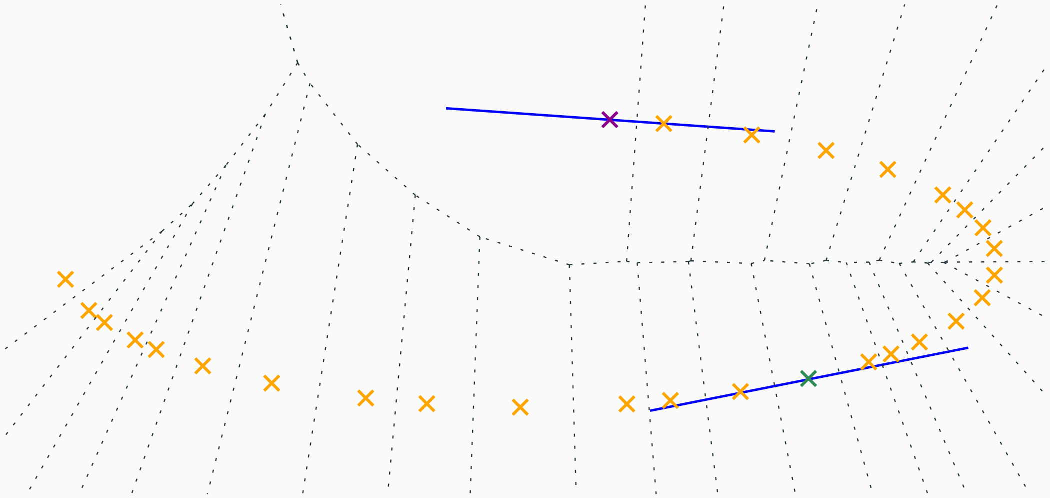
Insights on Boundary Point Detection



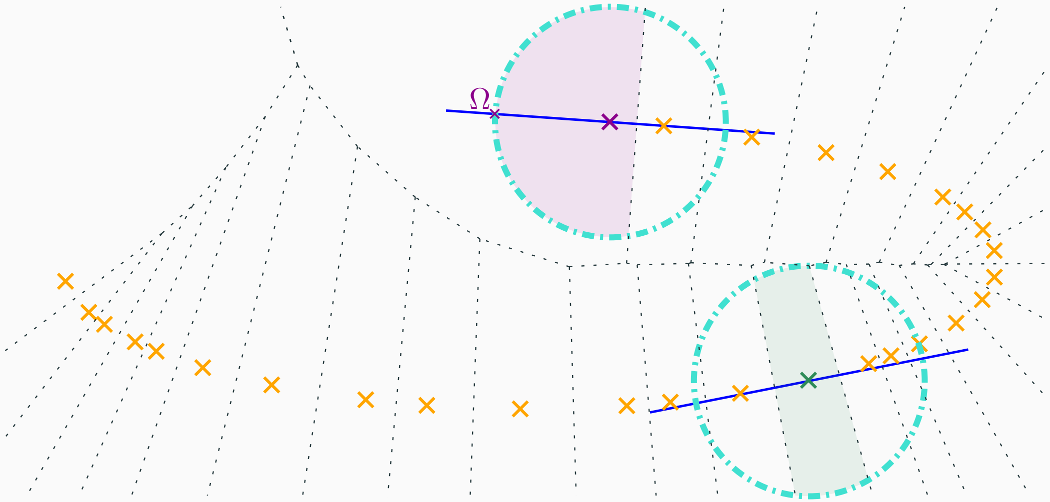
Insights on Boundary Point Detection



Insights on Boundary Point Detection



Insights on Boundary Point Detection



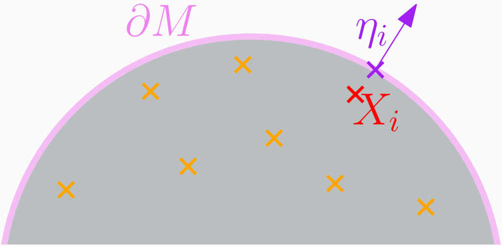
Boundary Observations: Definition

Write

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

Define the set of *boundary observations* as

$$\mathcal{Y}_{R_0,r,\rho} := \left\{ X_i \in \mathbb{X}_n \mid \exists X_j \in \mathrm{B}(X_i, r) \cap \mathbb{X}_n \text{s.t. } \mathrm{Diam}(\mathrm{Vor}_{R_0}^{(j)}(X_i)) \geq \rho \right\}.$$



Boundary Observations: Definition

Write

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

Define the set of *boundary observations* as

$$\mathcal{Y}_{R_0,r,\rho} := \left\{ X_i \in \mathbb{X}_n \mid \exists X_j \in \mathrm{B}(X_i, r) \cap \mathbb{X}_n \text{s.t. } \mathrm{Diam}(\mathrm{Vor}_{R_0}^{(j)}(X_i)) \geq \rho \right\}.$$

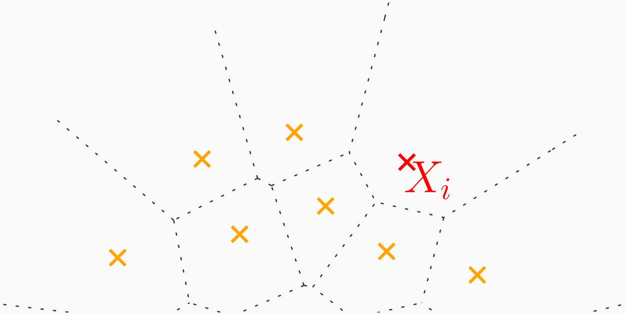
Boundary Observations: Definition

Write

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

Define the set of *boundary observations* as

$$\mathcal{Y}_{R_0,r,\rho} := \left\{ X_i \in \mathbb{X}_n \mid \exists X_j \in \mathrm{B}(X_i, r) \cap \mathbb{X}_n \text{s.t. } \mathrm{Diam}(\mathrm{Vor}^{(j)}_{R_0}(X_i)) \geq \rho \right\}.$$



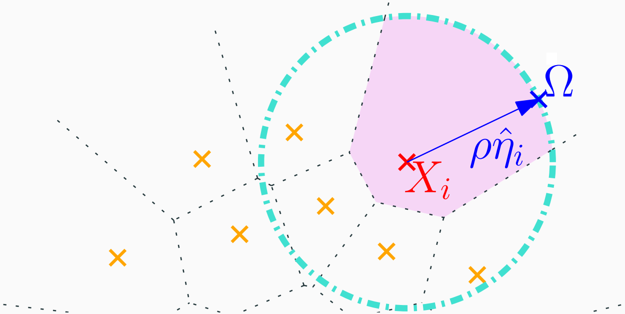
Boundary Observations: Definition

Write

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

Define the set of *boundary observations* as

$$\mathcal{Y}_{R_0,r,\rho} := \left\{ X_i \in \mathbb{X}_n \mid \exists X_j \in \mathrm{B}(X_i, r) \cap \mathbb{X}_n \text{s.t. } \mathrm{Diam}(\mathrm{Vor}^{(j)}_{R_0}(X_i)) \geq \rho \right\}.$$



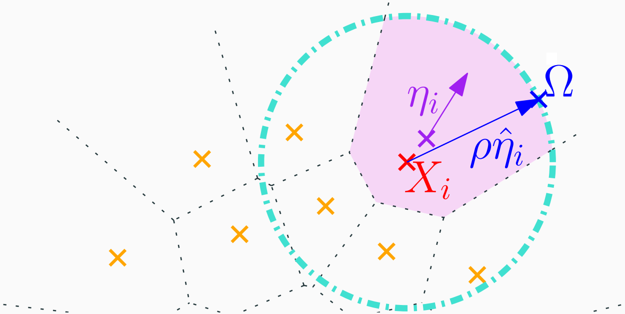
Boundary Observations: Definition

Write

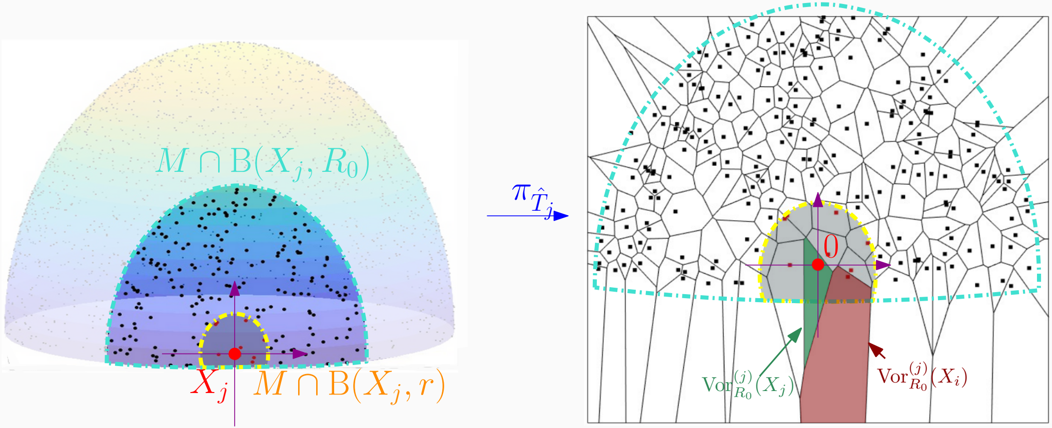
$$\mathrm{Vor}^{(j)}_{R_0}(X_i) := \left\{ O \in \hat{T}_j \,\middle|\, \mathring{\mathrm{B}}\big(O, \|O - \pi_{\hat{T}_j}(X_i - X_j)\|\big) \cap \pi_{\hat{T}_j}(\mathrm{B}(X_j, R_0) \cap \mathbb{X}_n - X_j) = \emptyset \right\}.$$

Define the set of *boundary observations* as

$$\mathcal{Y}_{R_0,r,\rho} := \left\{ X_i \in \mathbb{X}_n \mid \exists X_j \in \mathrm{B}(X_i, r) \cap \mathbb{X}_n \text{s.t. } \mathrm{Diam}(\mathrm{Vor}^{(j)}_{R_0}(X_i)) \geq \rho \right\}.$$



Boundary Observations: Illustration



Guarantees for Boundary Detection and Normals

Choosing the parameters properly, we have the following with high probability:

If $\partial M = \emptyset$, then $\mathcal{Y}_{R_0,r,\rho} = \emptyset$;
If $\partial M \neq \emptyset$ then:

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

$$\mathrm{d}(X_i, \partial M) \lesssim \left(\frac{\log n}{n}\right)^{\frac{2}{d+1}}.$$

For all $x \in \partial M$,

$$\mathrm{d}(x, \mathcal{Y}_{R_0,r,\rho}) \lesssim \left(\frac{\log n}{n}\right)^{\frac{1}{d+1}}.$$

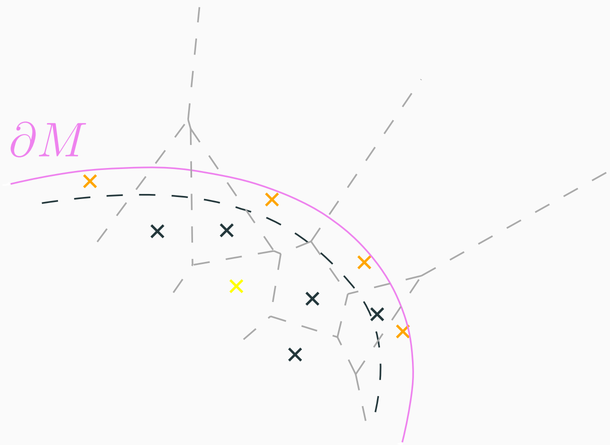
For all $X_i \in \mathcal{Y}_{R_0,r,\rho}$,

$$\|\eta_{\pi_{\partial M}(X_i)} - \tilde{\eta}_i\| \lesssim \left(\frac{\log n}{n}\right)^{\frac{1}{d+1}}.$$

Guarantees: Illustration

Write

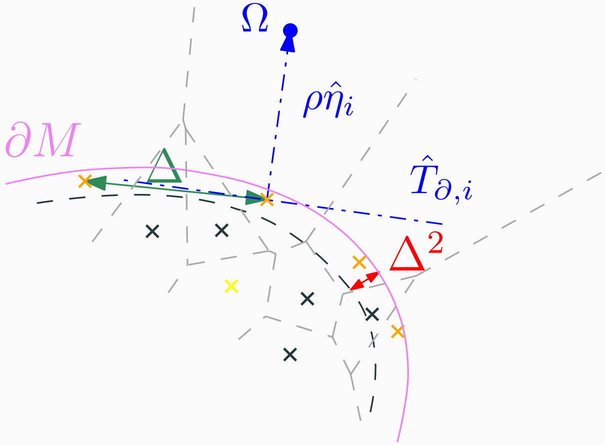
$$\Delta := \left(\frac{\log n}{n}\right)^{\frac{1}{d+1}}.$$



Guarantees: Illustration

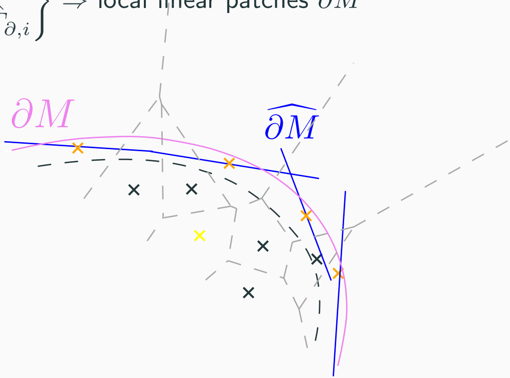
Write

$$\Delta := \left(\frac{\log n}{n}\right)^{\frac{1}{d+1}} .$$

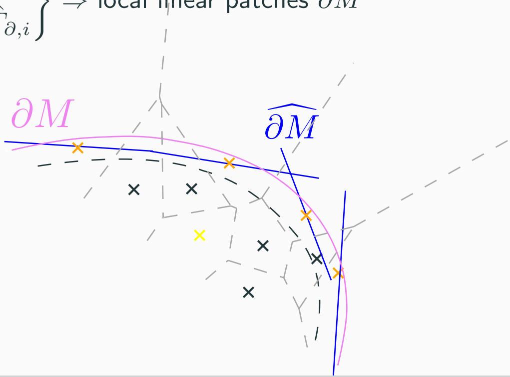


Boundary Estimation

$$\left.\begin{array}{l}\text{Boundary points } \mathcal{Y} \\ \text{Boundary's tangents estimates } \hat{T}_{\partial,i}\end{array}\right\} \Rightarrow \text{local linear patches } \widehat{\partial M}$$



$$\left.\begin{array}{l}\text{Boundary points } \mathcal{Y} \\ \text{Boundary's tangents estimates } \hat{T}_{\partial,i}\end{array}\right\} \Rightarrow \text{local linear patches } \widehat{\partial M}$$

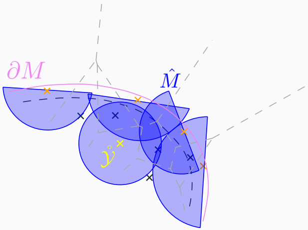


Theorem (Aamari, Aaron, Levrard – 2023)

$$\mathbb{E}[\mathrm{d_H}(\partial M, \widehat{\partial M})] \lesssim \left(\frac{\log n}{n}\right)^{\frac{2}{d+1}} \qquad \textit{(minimax optimal over } \mathcal{P}^{d,D}_{\mathrm{rch_{min}},\mathrm{rch}_{\partial,\mathrm{min}}}\textit{)}$$

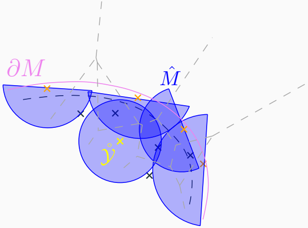
Estimation with Boundary

$$\left.\begin{array}{l}\text{Boundary \& Interior points } \mathcal{Y} \text{ \& } \mathbb{X}_n \setminus \mathcal{Y} \\ \text{Boundary's tangents estimates } \hat{T}_{\partial,i} \\ \text{Manifold's tangents estimates } \hat{T}_i\end{array}\right\} \Rightarrow \text{local linear patches / half-patches } \hat{M}$$



Estimation with Boundary

$$\left.\begin{array}{l}\text{Boundary \& Interior points } \mathcal{Y} \text{ \& } \mathbb{X}_n \setminus \mathcal{Y} \\ \text{Boundary's tangents estimates } \hat{T}_{\partial,i} \\ \text{Manifold's tangents estimates } \hat{T}_i\end{array}\right\} \Rightarrow \text{local linear patches / half-patches } \hat{M}$$

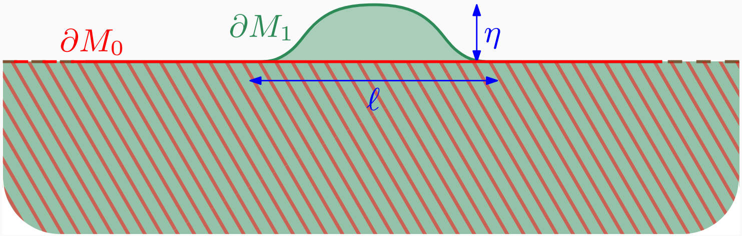


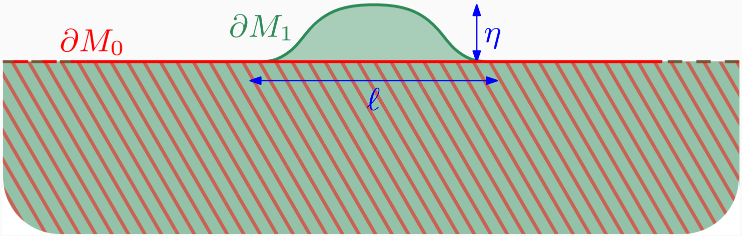
Theorem (Aamari, Aaron, Levrard – 2023)

$$\mathbb{E}[\mathrm{d_H}(M, \widehat{M})] \lesssim \left(\frac{\log n}{n}\right)^{\frac{2}{d+1}}$$

(minimax optimal over $\mathcal{P}^{d,D}_{\mathrm{rch}_{\min},\mathrm{rch}_{\partial,\min}}$*)*

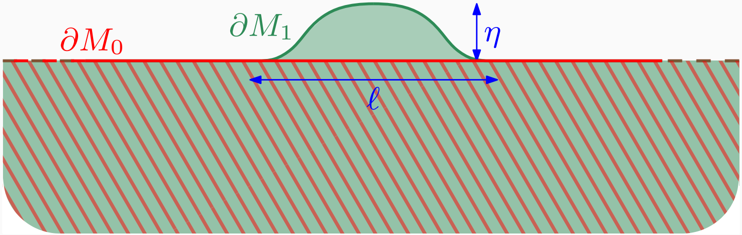
Le Cam's Lemma Heuristic





- P_0 and P_1 both belong to $\mathcal{P}^{\partial}$ as soon as $\eta \lesssim \ell^2$,
- $\mathrm{d_H}(M_0, M_1) \geq \eta$,
- $\mathrm{TV}(P_0, P_1) \lesssim \ell^{d-1}\eta$.

Le Cam's Lemma Heuristic

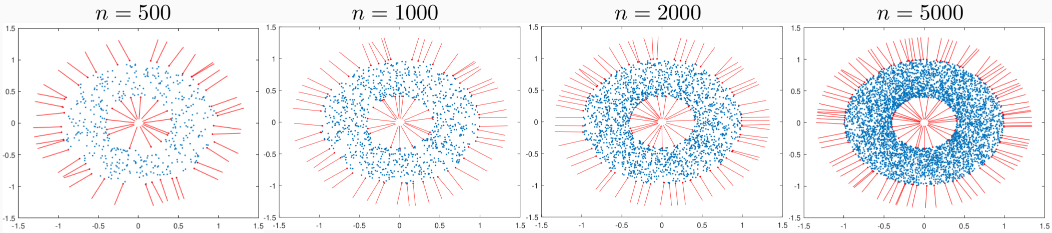


- P_0 and P_1 both belong to $\mathcal{P}^{\partial}$ as soon as $\eta \lesssim \ell^2$,
- $\mathrm{d_H}(M_0, M_1) \geq \eta$,
- $\mathrm{TV}(P_0, P_1) \lesssim \ell^{d-1}\eta$.

Hence, for $\eta \approx \ell^2$ and $\ell \approx (1/n)^{1/(d+1)}$,

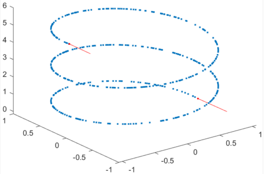
$$\inf_{\hat{M}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}\left[\mathrm{d_H}\left(M, \hat{M}_n\right)\right] \gtrsim \eta \left(1 - \ell^{d-1}\eta\right)^n \approx \ell^2 \left(1 - \ell^{d+1}\right)^n \approx (1/n)^{2/(d+1)}.$$

Annulus

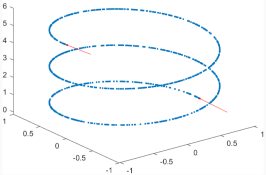


Spiral

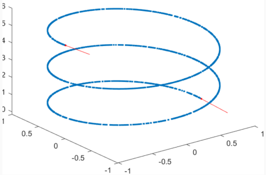
$n = 500$



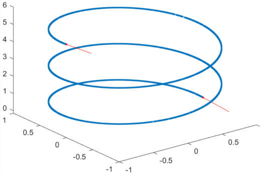
$n = 1000$

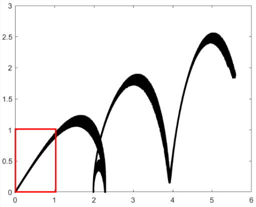


$n = 2000$

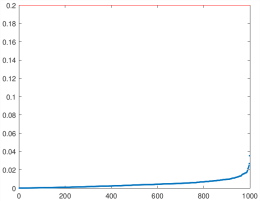


$n = 5000$





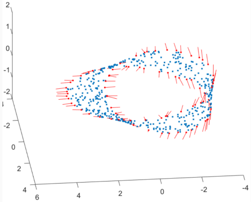
Calibration of R_0



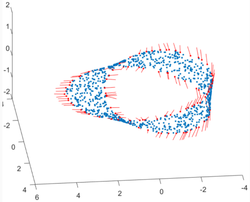
Calibration of ρ

Möbius strip

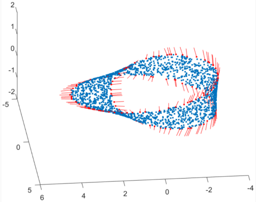
$n = 500$



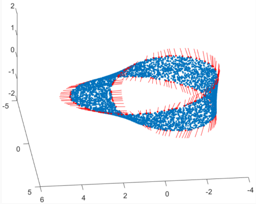
$n = 1000$

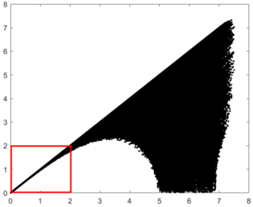


$n = 2000$

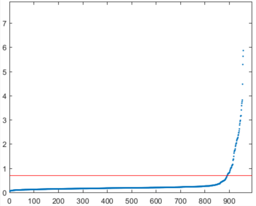


$n = 5000$





Calibration of R_0



Calibration of ρ

Influence of noise

Additive Noise

Crucial limitation: If significant noise is added, all the above methods fail!

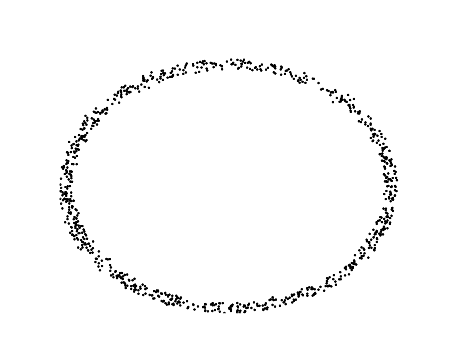


Figure 2: Circle with additive noise amplitude $\sigma > 0$

A Zoology of Noise Models

As opposed to nonparametric regression, many natural noise models:

- $Y = X + \varepsilon$ with $X \in M$
  and $X \perp\!\!\!\perp \varepsilon \in \mathbb{R}^D$ such that $\mathbb{E}[X] = 0$ (Convolution)
  [Fefferman et al. 2019]; [Genovese et al. 2012]
- $Y = X + \varepsilon$ with $X \in M$
  and $\varepsilon \in (T_X M)^\perp$ such that $\mathbb{E}[\varepsilon|X] = 0$ (Orthogonal noise)
  [Genovese et al. 2012b]; [Puchkin and Spokoiny 2022]
- $Y \sim \mathrm{Unif}_{M^\sigma}$ (Ambient uniform)
  [Aizenbud and Sober 2021]

A Zoology of Noise Models

As opposed to nonparametric regression, many natural noise models:

- $Y = X + \varepsilon$ with $X \in M$
  and $X \perp\!\!\!\perp \varepsilon \in \mathbb{R}^D$ such that $\mathbb{E}[X] = 0$ (Convolution)
  [Fefferman et al. 2019]; [Genovese et al. 2012]
- $Y = X + \varepsilon$ with $X \in M$
  and $\varepsilon \in (T_X M)^\perp$ such that $\mathbb{E}[\varepsilon | X] = 0$ (Orthogonal noise)
  [Genovese et al. 2012b]; [Puchkin and Spokoiny 2022]
- $Y \sim \mathrm{Unif}_{M^\sigma}$ (Ambient uniform)
  [Aizenbud and Sober 2021]

Take away:

Minimax rates for manifold estimation in the presence of noise are not fully understood.

About Noise not Being Centered

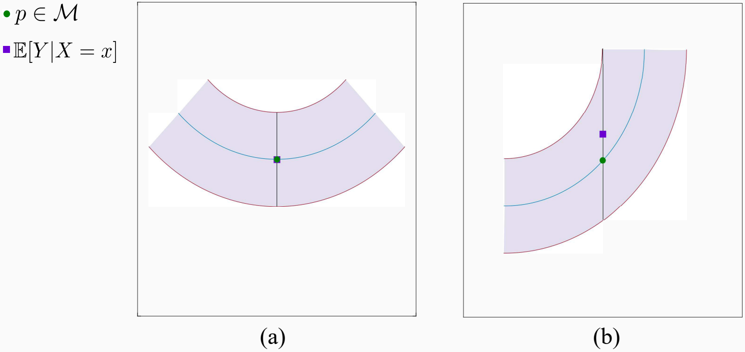


Figure 3: From [Aizenbud and Sober 2021]

Problem

An error in the tangent space yield apparent noise not centered, whatever the type of noise.

Alternating local PCA and Nonparametric Regression

Iterative algorithm [Puchkin and Spokoiny 2022] and [Aizenbud and Sober 2021]

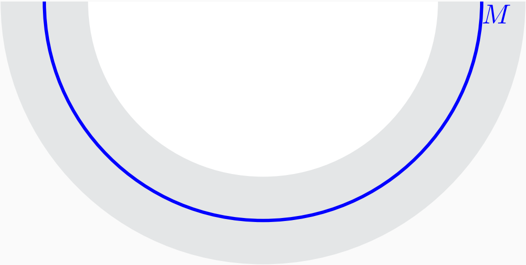
Tangents Initialize local coordinates with local PCA at scale $h_0 \simeq 1$.

Denoising In these coordinates, apply classical nonparametric regression at scale $h_1 < h_0$.

Tangents Store these new local coordinates and associated denoised points

Denoising In these coordinates, apply classical nonparametric regression at scale $h_2 < h_1$.

$\dots$



Alternating local PCA and Nonparametric Regression

Iterative algorithm [Puchkin and Spokoiny 2022] and [Aizenbud and Sober 2021]

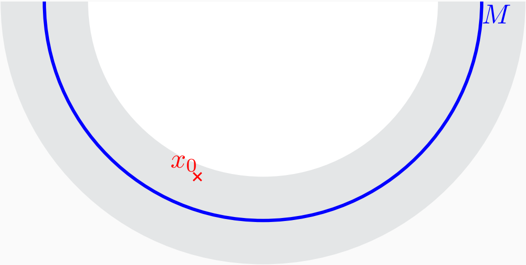
Tangents Initialize local coordinates with local PCA at scale $h_0 \simeq 1$.

Denoising In these coordinates, apply classical nonparametric regression at scale $h_1 < h_0$.

Tangents Store these new local coordinates and associated denoised points

Denoising In these coordinates, apply classical nonparametric regression at scale $h_2 < h_1$.

$\dots$



Alternating local PCA and Nonparametric Regression

Iterative algorithm [Puchkin and Spokoiny 2022] and [Aizenbud and Sober 2021]

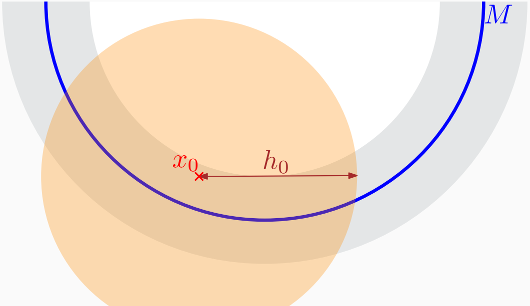
Tangents Initialize local coordinates with local PCA at scale $h_0 \simeq 1$.

Denoising In these coordinates, apply classical nonparametric regression at scale $h_1 < h_0$.

Tangents Store these new local coordinates and associated denoised points

Denoising In these coordinates, apply classical nonparametric regression at scale $h_2 < h_1$.

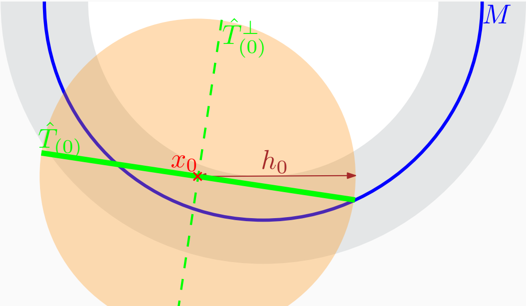
$\ldots$



Alternating local PCA and Nonparametric Regression

Iterative algorithm [Puchkin and Spokoiny 2022] and [Aizenbud and Sober 2021]

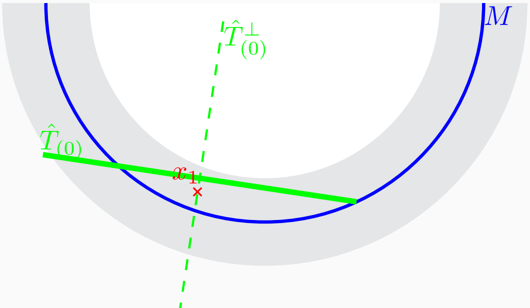
Tangents	Initialize local coordinates with local PCA at scale $h_0 \simeq 1$.
Denoising	In these coordinates, apply classical nonparametric regression at scale $h_1 < h_0$.
Tangents	Store these new local coordinates and associated denoised points
Denoising	In these coordinates, apply classical nonparametric regression at scale $h_2 < h_1$.
	...



Alternating local PCA and Nonparametric Regression

Iterative algorithm [Puchkin and Spokoiny 2022] and [Aizenbud and Sober 2021]

Tangents	Initialize local coordinates with local PCA at scale $h_0 \simeq 1$.
Denoising	In these coordinates, apply classical nonparametric regression at scale $h_1 < h_0$.
Tangents	Store these new local coordinates and associated denoised points
Denoising	In these coordinates, apply classical nonparametric regression at scale $h_2 < h_1$.
	...



Alternating local PCA and Nonparametric Regression

Iterative algorithm [Puchkin and Spokoiny 2022] and [Aizenbud and Sober 2021]

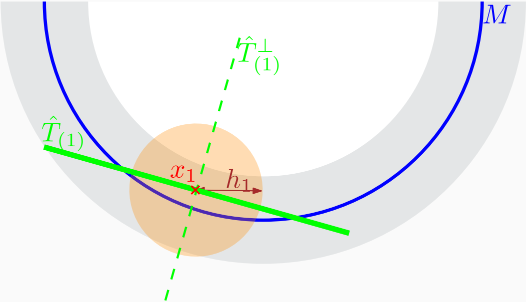
Tangents Initialize local coordinates with local PCA at scale $h_0 \simeq 1$.

Denoising In these coordinates, apply classical nonparametric regression at scale $h_1 < h_0$.

Tangents Store these new local coordinates and associated denoised points

Denoising In these coordinates, apply classical nonparametric regression at scale $h_2 < h_1$.

...



Alternating local PCA and Nonparametric Regression

Iterative algorithm [Puchkin and Spokoiny 2022] and [Aizenbud and Sober 2021]

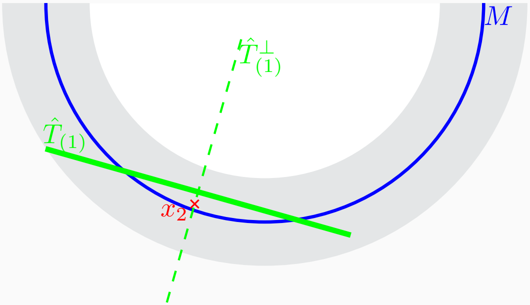
Tangents Initialize local coordinates with local PCA at scale $h_0 \simeq 1$.

Denoising In these coordinates, apply classical nonparametric regression at scale $h_1 < h_0$.

Tangents Store these new local coordinates and associated denoised points

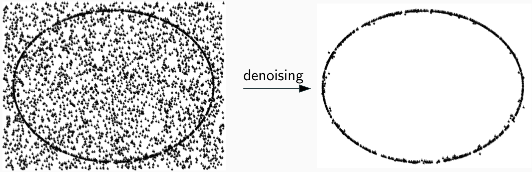
Denoising In these coordinates, apply classical nonparametric regression at scale $h_2 < h_1$.

$\dots$



Clutter Noise

$$\mathcal{P}^{(\mathrm{clutter})}_{\beta,Q_0} = \left\{\beta P + (1-\beta)Q_0, P \in \mathcal{P}\right\}.$$



Theorem (Aamari, Levrard – 2018)

With a decluttering procedure removing points from $Q_0 = \mathrm{Unif}_{\mathrm{B}(0,R)}$*, we can build an estimator such that*

$$\sup_{P \in \mathcal{P}^{(\mathrm{clutter})}} \mathbb{E}_{P^n}\left[\mathrm{d_H}\left(M, \hat{M}_n\right)\right] \lesssim \left(\frac{\log n}{n}\right)^{\frac{2}{d}}.$$

Remark: This procedure may fail for other Q_0's, even if Q_0 is known.

Parameter Selection

r-Convex Hull

For all $t \geq 0$, the t-convex hull of $A \subset \mathbb{R}^p$ is

$$\mathrm{Conv}_t(A) := \bigcup_{\substack{\sigma \subset A \\ \mathrm{rad}(\sigma) \leq t}} \mathrm{Conv}(A),$$

where $\mathrm{rad}(\sigma)$ is the radius of the smallest ball enclosing σ.

Figure 4: from Vicent Divol's PhD Defense

r-Convex Hull

For all $t \geq 0$, the t-convex hull of $A \subset \mathbb{R}^p$ is

$$\mathrm{Conv}_t(A) := \bigcup_{\substack{\sigma \subset A \\ \mathrm{rad}(\sigma) \leq t}} \mathrm{Conv}(A),$$

where $\mathrm{rad}(\sigma)$ is the radius of the smallest ball enclosing σ.

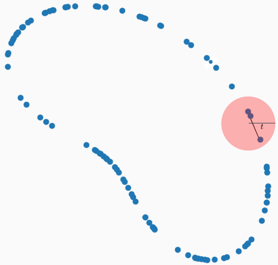


Figure 4: from Vicent Divol's PhD Defense

r-Convex Hull

For all $t \geq 0$, the t-convex hull of $A \subset \mathbb{R}^p$ is

$$\mathrm{Conv}_t(A) := \bigcup_{\substack{\sigma \subset A \\ \mathrm{rad}(\sigma) \leq t}} \mathrm{Conv}(A),$$

where $\mathrm{rad}(\sigma)$ is the radius of the smallest ball enclosing σ.

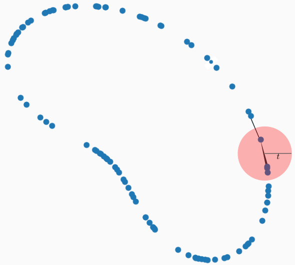


Figure 4: from Vicent Divol's PhD Defense

r-Convex Hull

For all $t \geq 0$, the t-convex hull of $A \subset \mathbb{R}^p$ is

$$\mathrm{Conv}_t(A) := \bigcup_{\substack{\sigma \subset A \\ \mathrm{rad}(\sigma) \leq t}} \mathrm{Conv}(A),$$

where $\mathrm{rad}(\sigma)$ is the radius of the smallest ball enclosing σ.

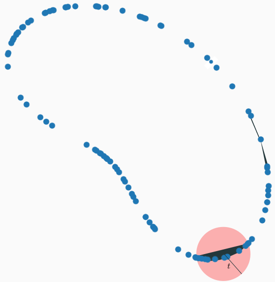


Figure 4: from Vicent Divol's PhD Defense

r-Convex Hull

For all $t \geq 0$, the t-convex hull of $A \subset \mathbb{R}^p$ is

$$\mathrm{Conv}_t(A) := \bigcup_{\substack{\sigma \subset A \\ \mathrm{rad}(\sigma) \leq t}} \mathrm{Conv}(A),$$

where $\mathrm{rad}(\sigma)$ is the radius of the smallest ball enclosing σ.

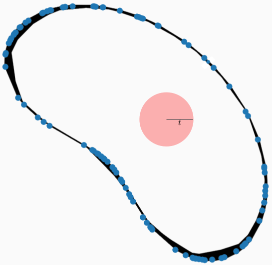


Figure 4: from Vicent Divol's PhD Defense

Reconstruction from r-Convex Hull

$$\text{Let } t^*(A) := \inf\left\{t < \mathrm{rch}_M \,\middle|\, \pi_M(\mathrm{Conv}_t(A)) = M\right\}$$

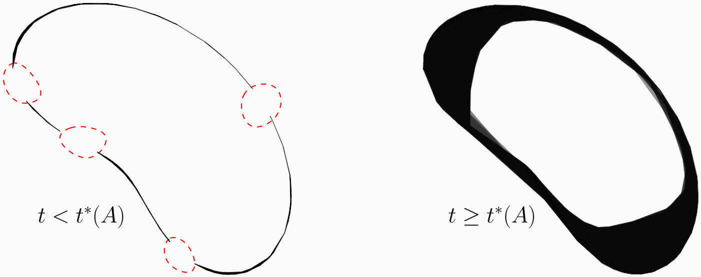


Figure 5: from Vicent Divol's PhD Defense

Reconstruction from r-Convex Hull

Let $t^*(A) := \inf\{t < \mathrm{rch}_M \,|\pi_M(\mathrm{Conv}_t(A)) = M\}$

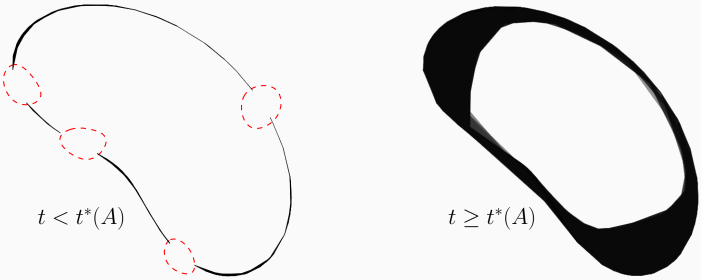


Figure 5: from Vicent Divol's PhD Defense

$\Rightarrow$ To reconstruct, need to pick $t > t^*(A)$ but as small as possible.

Reconstruction from r-Convex Hull

Theorem (divol2021minimax)

There exists $C = C_{\mathcal{P}} > 0$ such that picking $t = C\left(\log n/n\right)^{1/d}$, then for all $P \in \mathcal{P}$ and $n \geq 1$ large enough,

$$\mathrm{d_H}(M, \mathrm{Conv}_t(\mathcal{X}_n)) \lesssim \left(\frac{\log n}{n}\right)^{2/d}.$$

Reconstruction from r-Convex Hull

Theorem (divol2021minimax)

There exists $C = C_{\mathcal{P}} > 0$ such that picking $t = C\left(\log n/n\right)^{1/d}$, then for all $P \in \mathcal{P}$ and $n \geq 1$ large enough,

$$\mathrm{d_H}(M, \mathrm{Conv}_t(\mathcal{X}_n)) \lesssim \left(\frac{\log n}{n}\right)^{2/d}.$$

Limitation

In practice, need to calibrate the constant C.

(or equivalently t)

Idea

Compare each estimator $\mathrm{Conv}_t(\mathcal{X}_n)$ with the most overfitting one $\mathrm{Conv}_t(\mathcal{X}_n) = \mathcal{X}_n$ of the family.

Convexity Defect Function

The convexity defect function of $A \subset \mathbb{R}^p$ at scale $t \geq 0$ is

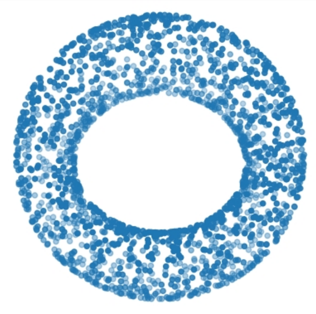
$$h(t, A) := \mathrm{d_H}(A, \mathrm{Conv}_t(A))$$

Convexity Defect Function

The convexity defect function of $A \subset \mathbb{R}^p$ at scale $t \geq 0$ is

$$h(t, A) := \mathrm{d_H}(A, \mathrm{Conv}_t(A))$$

If $\mathrm{rch}_M > 0$, then $h(t, M) \leq t^2/\mathrm{rch}_M$
For point clouds $A = \mathcal{X}_n$, the behavior looks like this:



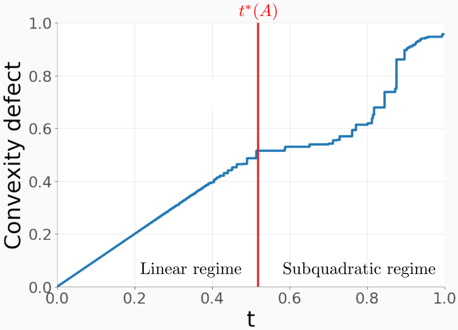


Figure 6: from Vicent Divol's PhD Defense

Scale Parameter Choice

Given $0 < \lambda \leq 1$, define

$$t_\lambda(A) := \inf\{t \in \mathrm{Rad}(A) | h(t, A) \leq \lambda t\},$$

where $\mathrm{Rad}(A) = \{\mathrm{rad}(\sigma)\}_{\sigma \subset A}$.

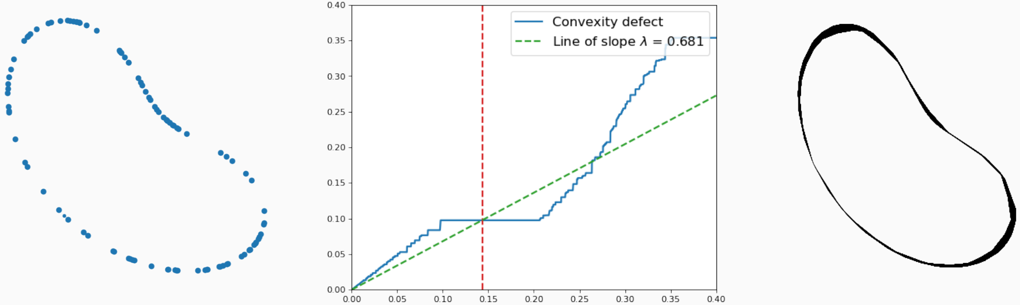


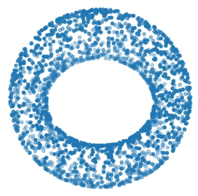
Figure 7: from Vicent Divol's PhD Defense

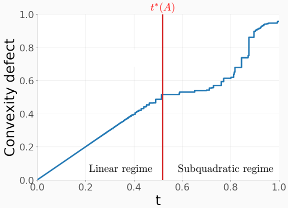
Scale-Free Manifold Estimation

Theorem (divol2021minimax)

Uniformly over $\mathcal{P}$, for all $n \geq 1$ large enough,

$$\mathbb{E}_{P^n}\left[\mathrm{d_H}\left(M, \mathrm{Conv}_{t_\lambda(\mathcal{X}_n)}(\mathcal{X}_n)\right)\right] \lesssim \left(\frac{\log n}{n}\right)^{2/d}.$$

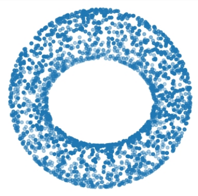


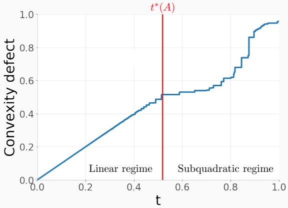


Theorem (divol2021minimax)

Uniformly over $\mathcal{P}$, for all $n \geq 1$ large enough,

$$\mathbb{E}_{P^n}\left[\mathrm{d_H}\left(M, \mathrm{Conv}_{t_\lambda(\mathcal{X}_n)}(\mathcal{X}_n)\right)\right] \lesssim \left(\frac{\log n}{n}\right)^{2/d}.$$





Remark: This method is *not* fully parameter-free: choice of $\lambda \geq 1$.

Yet, $\lambda = 1/\sqrt{2}$ works (theoretically) for any dimension $d \geq 1$.

Smoother Manifolds

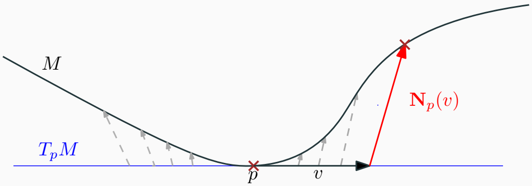
More Regularity

Definition ($\mathcal{C}^2$ Regularity Class)

Submanifolds $M \in \mathcal{C}^2_{\mathrm{rch}_{\min}}$ have local parametrizations

$$\begin{aligned} \Psi_p : T_pM &\longrightarrow M \subset \mathbb{R}^p \\ v &\longmapsto p + v + \mathbf{N}_p(v) \end{aligned}$$

where $\mathbf{N}_p(0) = 0$, $d_0\mathbf{N}_p = 0$ and $\|d_v\mathbf{N}_p\|_{op} \leq \|v\|/(2\mathrm{rch}_{\min})$.

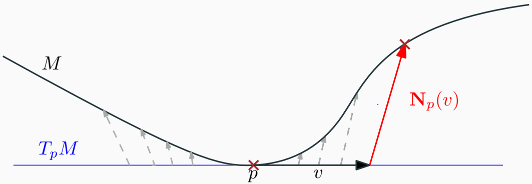


Definition ($\mathcal{C}^k$ Regularity Class, $k \geq 3$)

Let $\mathbf{L} = (L_2, L_3, \ldots, L_k)$, and define $\mathcal{C}^k_{\mathrm{rch_{min}},\mathbf{L}}$ to be the subset of elements $M \in \mathcal{C}^2_{\mathrm{rch_{min}}}$ that have local parametrizations

$$\begin{aligned} \Psi_p : T_pM &\longrightarrow M \subset \mathbb{R}^p \\ v &\longmapsto p + v + \mathbf{N}_p(v) \end{aligned}$$

where $\mathbf{N}_p(0) = 0$, $d_0\mathbf{N}_p = 0$ and $\|d_v^i \mathbf{N}_p\|_{op} \leq L_i$ for $2 \leq i \leq k$.



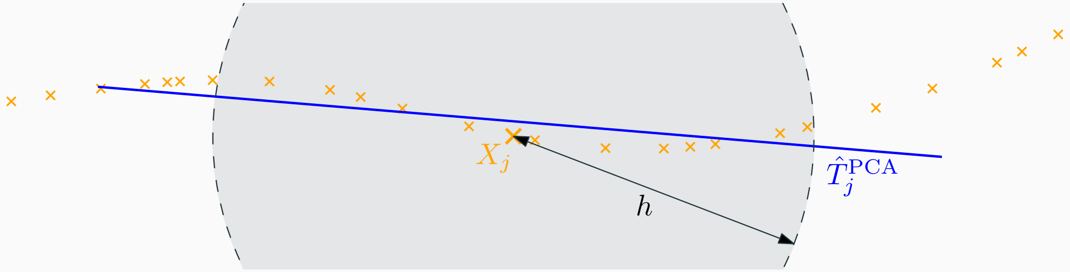
Local PCA

Recall that $P_n^{(j)} = \frac{1}{n}\sum_{\ell \neq j} \delta_{X_\ell - X_j}$, and

$$\hat{T}_j^{\texttt{PCA}} \in \underset{T \in \mathbb{G}^{D,d}}{\arg\min}\, P_n^{(j)} \left[\|x - \pi_T(x)\|^2 \mathbb{I}\{B(0,h)\}(x) \right].$$

$\mathbb{G}^{p,d}$: space of d-dimensional linear subspaces of $\mathbb{R}^p$;

π_T : orthogonal projection onto T.



Local Polynomials

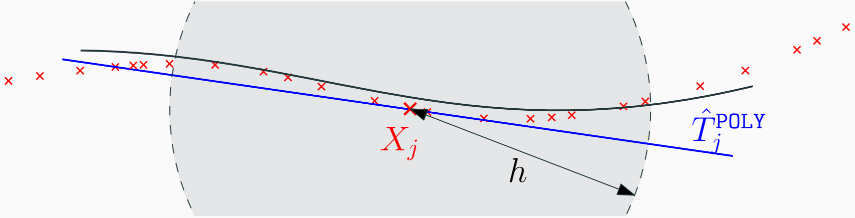
Define $\left(\hat{T}_j^{\mathtt{POLY}}, \hat{T}_{2,j}, \ldots, \hat{T}_{k-1,j}\right)$ to be a minimizer of

$$P_n^{(j)}\left[\left\|x - \pi_T(x) - \sum_{i=2}^{k-1} T^{(i)}\left(\pi_T(x)^{\otimes i}\right)\right\|^2 \mathbb{I}\{B(0,h)\}(x)\right],$$

where

T: ranges in $\mathbb{G}^{p,d}$;

$T^{(i)}$: ranges in the set of i-linear maps $(2 \leq i \leq k-1)$.



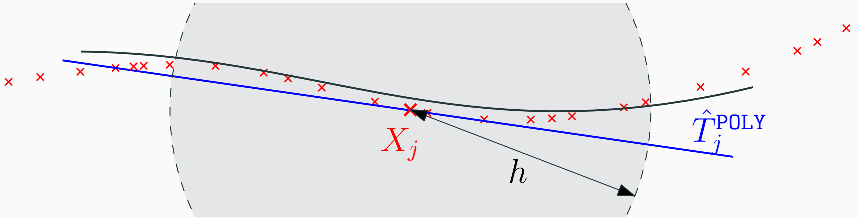
Similar methods in **Cazals06**; **Cheng16**; **sober2020manifold**.

Convergence of Local Polynomials

Theorem (Aamari19b)

If $h = C\left(\frac{\log n}{n}\right)^{1/d}$*, for all* $P \in \mathcal{P}^k_{\mathrm{rch}_{\min},\mathbf{L}}$*,*

$$\mathbb{E}_{P^n}\, \mathrm{d_H}\big(M, \hat{M}_{\texttt{POLY}}\big) \lesssim \left(\frac{\log n}{n}\right)^{\frac{k}{d}}.$$



↪ This rate is minimax optimal.

↪ Estimation of tangent spaces and curvature in the process