Understanding the geometry of high-dimensional data through the reach

GESDA 2022 — October 4, 2022

Clément Berenfeld (CEREMADE)

Joint work with **Eddie Aamari** and **Clément Levrard** (LPSM)

Introduction

High-dim. data $\rightarrow$ hidden low-dim. geometric structures

- Physical constraints;
- Implicit parametrisations.

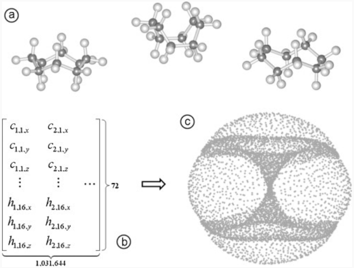

Figure 1: Cyclooctane conformations (Martin et al., 2010).

Introduction

High-dim. data $\rightarrow$ hidden low-dim. geometric structures

- Physical constraints;
- Implicit parametrisations.

Figure 1: Data from the Coil-20 dataset (Nene et al., 1996).

Introduction

High-dim. data → hidden low-dim. geometric structures

- ▷ Physical constraints;
- ▷ Implicit parametrisations.

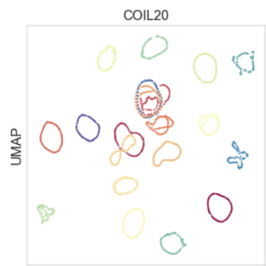

Figure 1: UMAP representation of Coil-20 (McInnes et al., 2018).

Introduction

We observe n points $X_1, \ldots, X_n$ lying on an unknown **submanifold**.

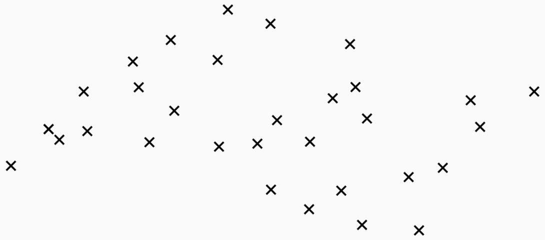

Figure 2: A point cloud and some interpolating shapes.

Introduction

We observe n points $X_1, \ldots, X_n$ lying on an unknown **submanifold**.

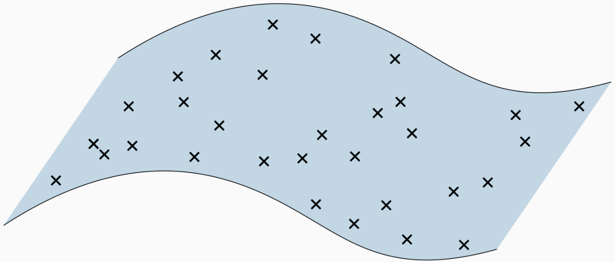

Figure 2: A point cloud and some interpolating shapes.

Introduction

We observe n points $X_1, \ldots, X_n$ lying on an unknown **submanifold**.

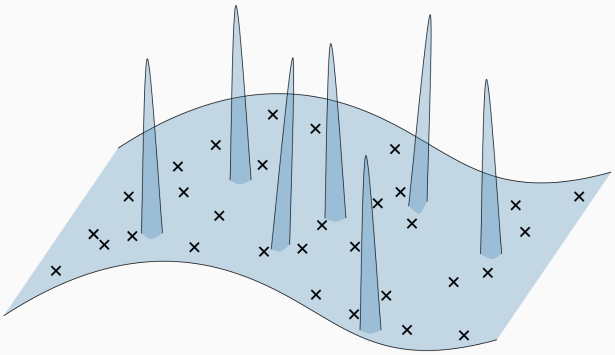

Figure 2: A point cloud and some interpolating shapes.

Introduction

We observe n points $X_1, \ldots, X_n$ lying on an unknown **submanifold**.

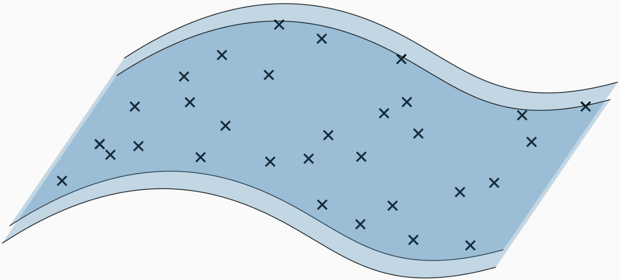

Figure 2: A point cloud and some interpolating shapes.

Introduction

We observe n points $X_1, \ldots, X_n$ lying on an unknown **submanifold**.

Figure 2: A point cloud and some interpolating shapes.

Introduction

All these shapes have a very different *resolution*:

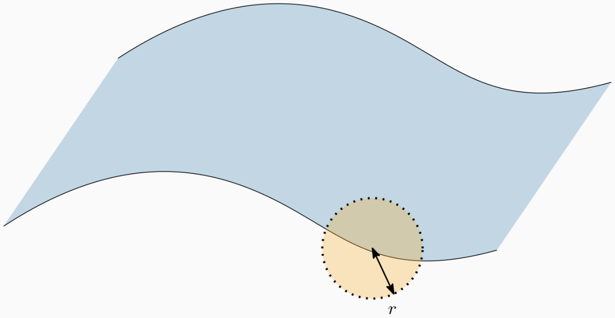

Figure 3: The behavior of the interpolating shapes at some scale $r > 0$.

All these shapes have a very different *resolution*:

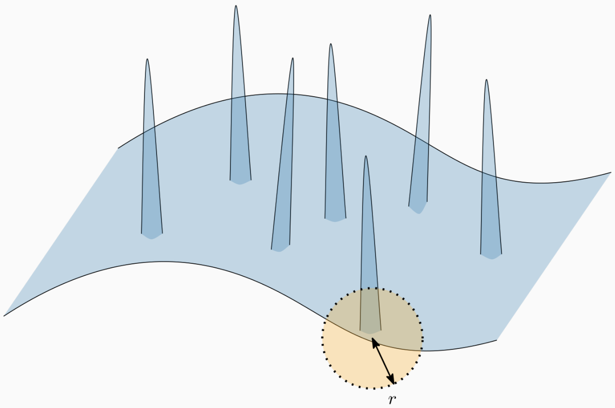

Figure 3: The behavior of the interpolating shapes at some scale $r > 0$.

All these shapes have a very different *resolution*:

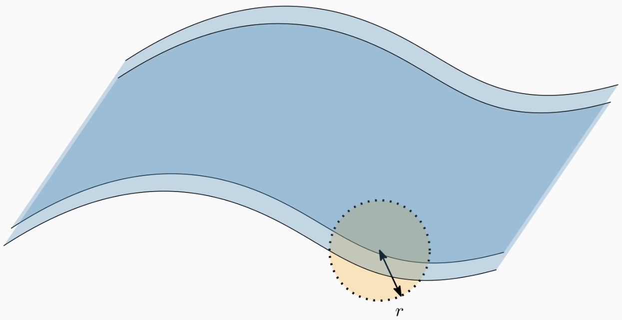

Figure 3: The behavior of the interpolating shapes at some scale $r > 0$.

All these shapes have a very different *resolution*:

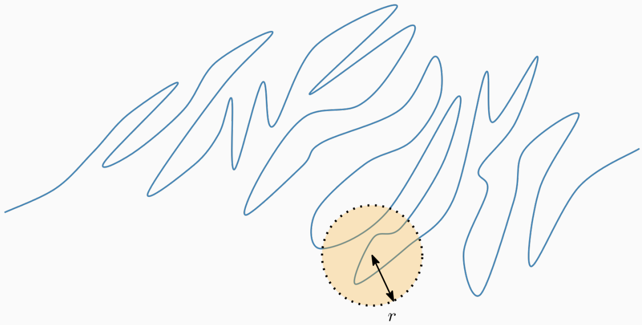

Figure 3: The behavior of the interpolating shapes at some scale $r > 0$.

Introduction

This resolution is called the ***reach*** of the support.

- ▷ It appears as a parameter in most statistical procedures;
- ▷ It drives the performance of estimators;
- ▷ Inference is mostly impossible without constraining it;

Goal: estimate the reach of the support of the underlying probability distribution.

Introduction

What has been done so far:

- Reach estimation on $\mathscr{C}^3$-model (Aamari et al., 2019);
- Optimal reach estimation up to the regularity $\mathscr{C}^4$ (B., Harvey, Hoffmann & Shankar, 2022);
- Universally consistent estimation of the reach (Cholaquidis et al., 2021).

What we will do today:

- Optimal reach estimation on $\mathscr{C}^k$ model;
- Optimal estimation of other scales along the way;
- Optimal metric learning.

Introduction

What has been done so far:

- Reach estimation on $\mathscr{C}^3$-model (Aamari et al., 2019);
- Optimal reach estimation up to the regularity $\mathscr{C}^4$ (B., Harvey, Hoffmann & Shankar, 2022);
- Universally consistent estimation of the reach (Cholaquidis et al., 2021).

What we will do today:

- Optimal reach estimation on $\mathscr{C}^k$ model;
- Optimal estimation of other scales along the way;
- Optimal metric learning.

The reach: definition and model

Definition of the reach

The reach (Federer, 1959) of $K \subset \mathbb{R}^D$ is defined as

$$\mathrm{rch}(K) := \sup\left\{r \geqslant 0 \mid \forall x \in K^r,\ \exists! y \in K,\ d(x,K) = \|x - y\|\right\}.$$

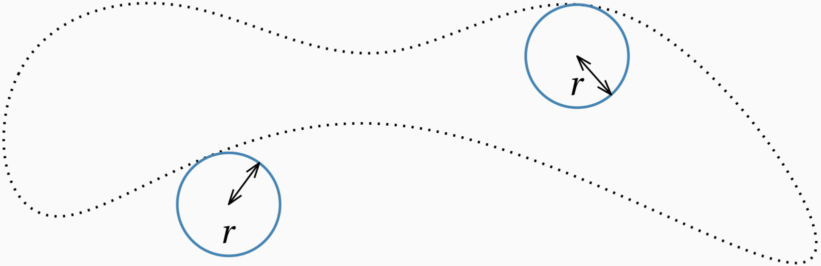

Figure 4: The rolling ball condition.

Definition of the reach

A reach constraint tend to discard support that are either too curved or too close to self-intersect.

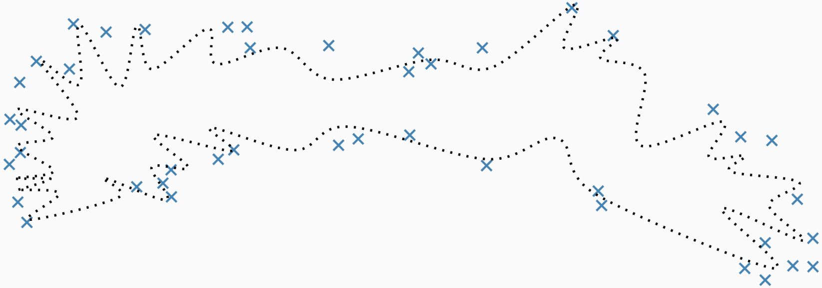

Figure 5: Quasi-interpolating shape for the point cloud that **does not meet** a reach constraint.

Definition of the reach

A reach constraint tend to discard support that are either too curved or too close to self-intersect.

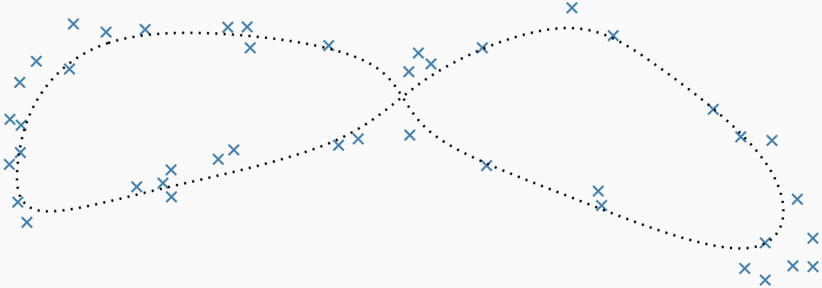

Figure 5: Quasi-interpolating shape for the point cloud that **does not meet** a reach constraint.

Definition of the reach

A reach constraint tend to discard support that are either too curved or too close to self-intersect.

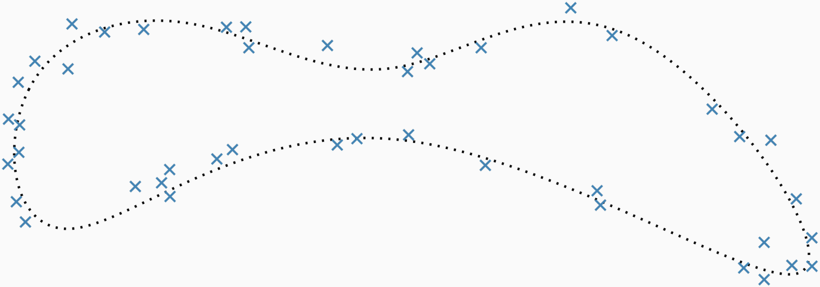

Figure 5: Quasi-interpolating shape for the point cloud that **meets** a reach constraint.

Statistical model

Let Σ_k be the set of probability measures P on $\mathbb{R}^D$ such that

1. P is supported on M, a d-dimensional, compact and $\mathscr{C}^k$ submanifold of $\mathbb{R}^D$;
2. The reach of M is lower-bounded by $\tau > 0$;
3. The density of P wrt to $\mathscr{H}^d|_M$ is bounded from above and below.

A key result

Hausdorff distance between two subsets $A, B \subset \mathbb{R}^D$:

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

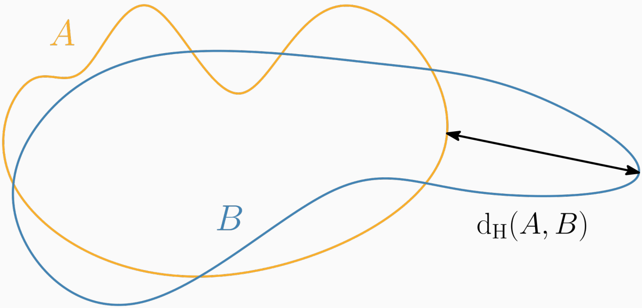

Figure 6: The Hausdorff distance between A and B.

A key result

Hausdorff distance between two subsets $A, B \subset \mathbb{R}^D$:

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

Theorem Aamari & Levrard (2019)

There exists an estimator $\widehat{M}$ such that such that for any $k \geqslant 3$,

$$\sup_{P\in\Sigma_k} \mathbb{E}_{P^{\otimes n}}[\mathrm{d_H}(\widehat{M}, M))] \preccurlyeq n^{-k/d},$$

and $\widehat{M}$ is obtained through local polynomial patching.

As a comparison:

- $\widehat{M} = \{X_1, \ldots, X_n\}$ has a risk of $n^{-1/d}$;
- $\widehat{M} =$ good triangulation has a risk of $n^{-2/d}$.

A key result

Remark: The reach is Hausdorff **unstable**.

Figure 6: A small Hausdorff perturbation, a significant change in reach.

- In particular, $\mathrm{rch}(\widehat{M}) \approx 0$ most of the time.
- Naive plug-in won't work.

A key result

Remark: The reach is Hausdorff **unstable**.

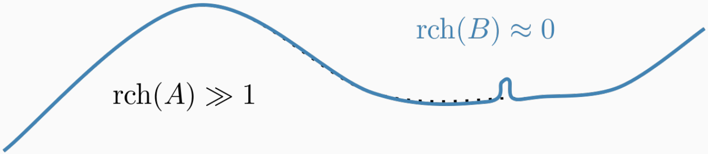

Figure 6: A small Hausdorff perturbation, a significant change in reach.

- In particular, $\mathrm{rch}(\widehat{M}) \approx 0$ most of the time.
- Naive plug-in won't work.

A key result

Remark: The reach is Hausdorff **unstable**.

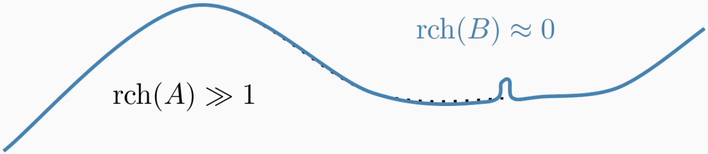

Figure 6: A small Hausdorff perturbation, a significant change in reach.

- In particular, $\mathrm{rch}(\widehat{M}) \approx 0$ most of the time.
- Naive plug-in won't work.

Estimation strategies for the reach

Reach decomposition

General idea: Leverage the decomposition result of

Theorem Aamari et al. (2019)

For any submanifold M, there holds

$$\mathrm{rch}(M) := R_\ell(M) \wedge \mathrm{wfs}(M).$$

The **local reach** $R_\ell(M)$ is the minimal radius of curvature of M

$$R_\ell(M) := \inf_{x \in M} \| \mathrm{II}_x \|_{\mathrm{op}}^{-1}.$$

The **weak feature size** $\mathrm{wfs}(M)$ is an important topological scale introduced by (Chazal and Lieutier, 2004).

$$\mathrm{wfs}(M) := \inf\left\{ r \geqslant 0 \mid \exists x \in M^r \setminus M,\ x \in \mathrm{Conv}\left(\mathrm{pr}_M(x)\right)\right\}.$$

Reach decomposition

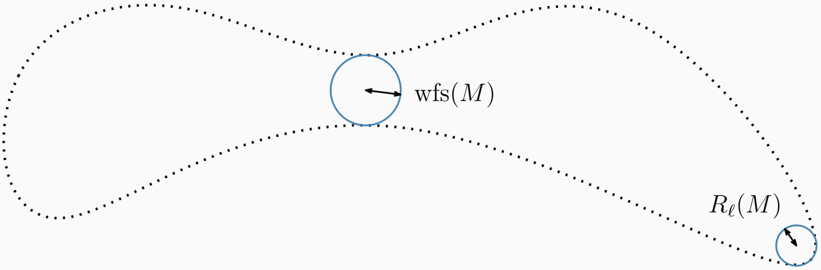

Figure 7: The weak feature size and local reach of a submanifold.

First step: Estimate $R_\ell(M)$.

- Compute the curvatures of a locally smooth support estimator.

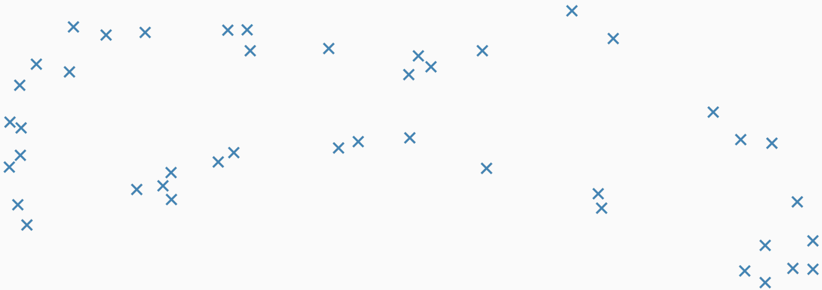

Figure 8: Estimating the local reach.

Curvature estimation

First step: Estimate $R_\ell(M)$.

- ▹ Compute the curvatures of a locally smooth support estimator.

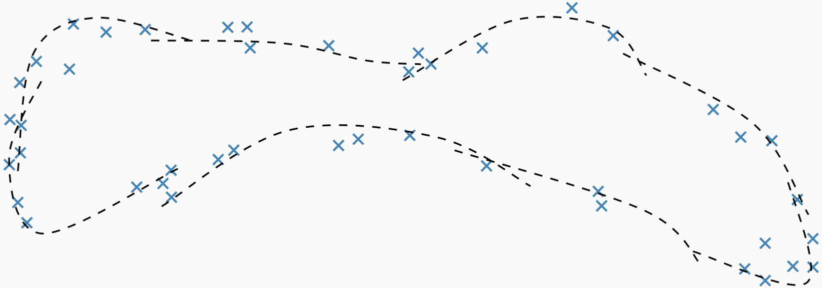

Figure 8: Estimating the local reach.

Curvature estimation

First step: Estimate $R_\ell(M)$.

- Compute the curvatures of a locally smooth support estimator.

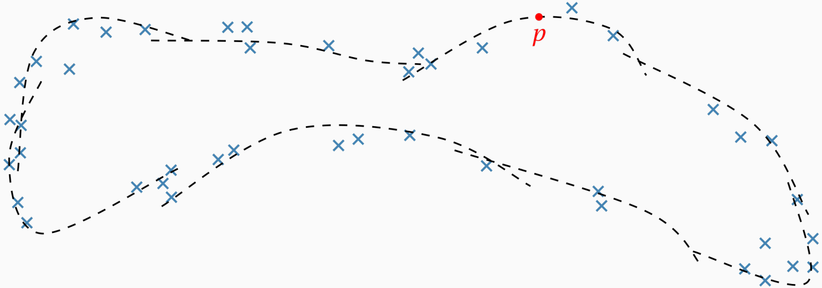

Figure 8: Estimating the local reach.

First step: Estimate $R_\ell(M)$.

- Compute the curvatures of a locally smooth support estimator.

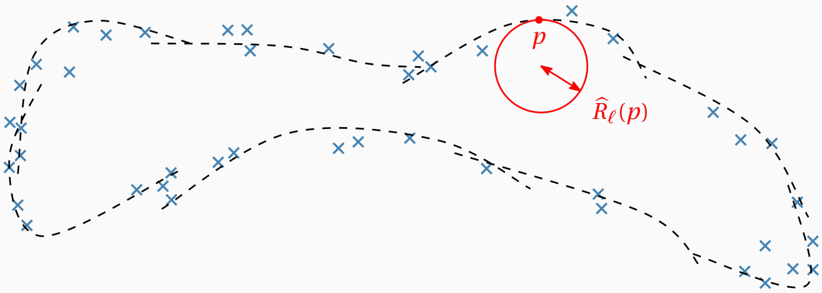

Figure 8: Estimating the local reach.

When applied to a polynomial patching of order k (Aamari and Levrard, 2019), the resulting estimator satisfies:

Theorem Aamari, B. & Levrard (2022)

For any $k \geqslant 3$,

$$\sup_{P \in \Sigma_k} \mathbb{E}_{P^{\otimes n}} \left[|\widehat{R_\ell} - R_\ell(M)| \right] \preccurlyeq n^{-\frac{k-2}{d}},$$

and this rate is minimax-optimal.

Finding a global scale

Next step: estimate $\mathrm{wfs}(M)$.

Theorem Aamari, B. & Levrard (2022)

For any $k \geqslant 3$,

$$\inf_{\tilde{w}} \sup_{P \in \Sigma_k} \mathbb{E}_{P^{\otimes n}} [|\tilde{w} - \mathrm{wfs}(M)|] \geqslant r_* > 0 \qquad \forall n \geqslant 1.$$

Idea: For any other interpolating scale $\mathrm{rch}(M) \leqslant \theta(M) \leqslant \mathrm{wfs}(M)$,

$$\mathrm{rch}(M) = R_\ell(M) \wedge \theta(M).$$

- Example of such scale: the μ-reach (Chazal et al., 2006).

Finding a global scale

Next step: estimate $\mathrm{wfs}(M)$.

Theorem Aamari, B. & Levrard (2022)

For any $k \geqslant 3$,

$$\inf_{\widetilde{w}} \sup_{P \in \Sigma_k} \mathbb{E}_{P^{\otimes n}} [|\widetilde{w} - \mathrm{wfs}(M)|] \geqslant r_* > 0 \qquad \forall n \geqslant 1.$$

Idea: For any other interpolating scale $\mathrm{rch}(M) \leqslant \theta(M) \leqslant \mathrm{wfs}(M)$,

$$\mathrm{rch}(M) = R_\ell(M) \wedge \theta(M).$$

- Example of such scale: the μ-reach (Chazal et al., 2006).

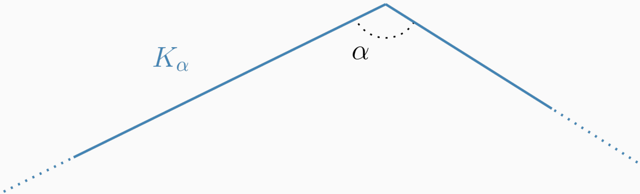

Figure 9: Two half-lines meeting with an angle $\alpha \in (0, \pi]$.

Finding a global scale

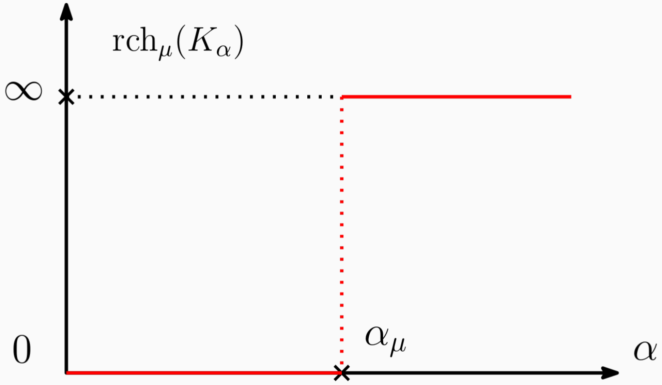

Figure 10: The instability of the μ-reach in a simple example.

Finding a global scale

Idea: Leverage the result

Theorem Boissonnat, Lieutier & Wintraecken (2019)

For any submanifold M,

$$\mathrm{rch}(M) = \sup\left\{r \mid \forall x, y \in M,\ \|x-y\| \leqslant 2r \Rightarrow\ \mathrm{d}_M(x,y) \leqslant \mathrm{d}_{\mathscr{S}(r)}(x,y)\right\}$$

where $\mathrm{d}_{\mathscr{S}(r)}(x,y) := 2r\arcsin\left(\frac{\|x-y\|}{2r}\right)$ is the spherical distance.

- ▷ Estimating the reach of M boils down to comparing the intrinsic distance on M with the spherical distances.

Spherical distortion radius

We define for any $\Delta > 0$,

$$\mathrm{sdr}_\Delta(M) := \sup\left\{r \mid \forall x, y \in M,\ \Delta \leqslant \|x - y\| \leqslant 2r \Rightarrow \mathrm{d}_M(x,y) \leqslant \mathrm{d}_{\mathscr{S}(r)}(x,y)\right\}.$$

Theorem Aamari, B. & Levrard (2022)

There holds, for any $0 \leqslant \Delta \leqslant \sqrt{\frac{2(D+1)}{D}}\,\mathrm{wfs}(M)$,

$$\mathrm{rch}(M) \leqslant \mathrm{sdr}_\Delta(M) \leqslant \mathrm{wfs}(M).$$

- ▹ To estimate $\mathrm{sdr}_\Delta(M)$, one can estimate M and d_M.
- ▹ Need to ensure stability of sdr with respect to (M, d_M).

Spherical distortion radius

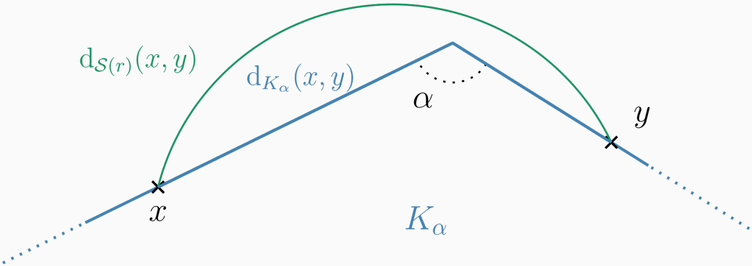

Figure 11: The stability of the sdr in a simple example.

Spherical distortion radius

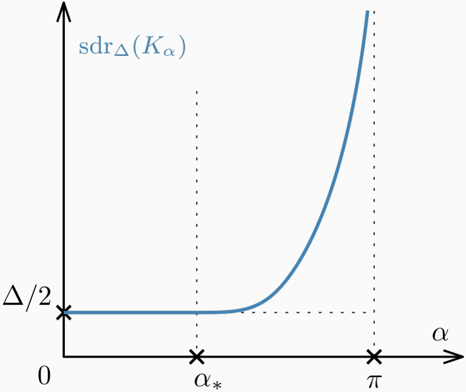

Figure 11: The stability of the sdr in a simple example.

Optimal metric learning

Isomap (Bernstein et al., 2000):

1. From the point cloud, build a neighborhood graph $\widehat{G}$;
2. Estimate $\widehat{\mathrm{d}}(x, y) := \mathrm{d}_{\widehat{G}}(x, y)$.

For a wisely chosen connectivity radius, one can get

$$(1 - \varepsilon_n)\widehat{\mathrm{d}}(x, y) \leqslant \mathrm{d}_M(x, y) \leqslant (1 + \varepsilon_n)\widehat{\mathrm{d}}(x, y),$$

with high probability and with $\varepsilon_n \approx n^{-2/3d}$, as shown in (Aaron & Bodart, 2018).

▷ There is room for improvement.

Isomap (Bernstein et al., 2000):

1. From the point cloud, build a neighborhood graph $\widehat{G}$;
2. Estimate $\widehat{\mathrm{d}}(x,y) := \mathrm{d}_{\widehat{G}}(x,y)$.

For a wisely chosen connectivity radius, one can get

$$(1-\varepsilon_n)\widehat{\mathrm{d}}(x,y) \leqslant \mathrm{d}_M(x,y) \leqslant (1+\varepsilon_n)\widehat{\mathrm{d}}(x,y),$$

with high probability and with $\varepsilon_n \approx n^{-2/3d}$, as shown in (Aaron & Bodart, 2018).

- ▷ There is room for improvement.

Metric learning

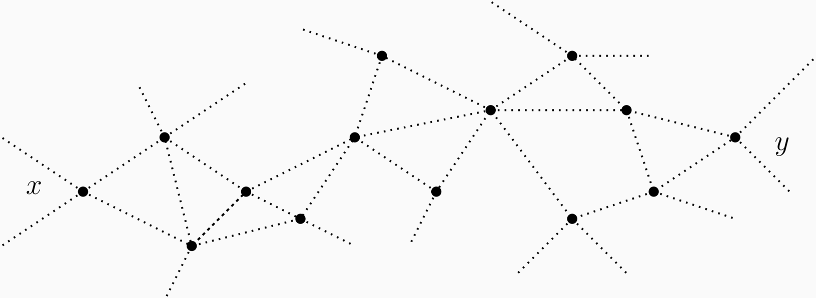

Figure 12: Enhancing the Isomap algorithm (Aamari, B. & Levard, 2022).

▷ The accuracy becomes $\varepsilon_n \approx n^{-1/d}$: that's better.

Metric learning

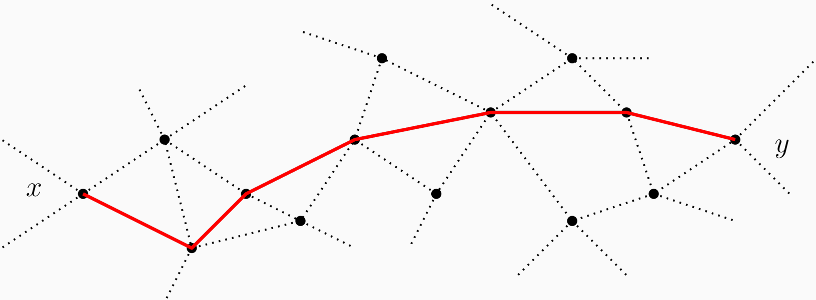

Figure 12: Enhancing the Isomap algorithm (Aamari, B. & Levard, 2022).

- The accuracy becomes $\varepsilon_n \approx n^{-1/d}$: that's better.

Metric learning

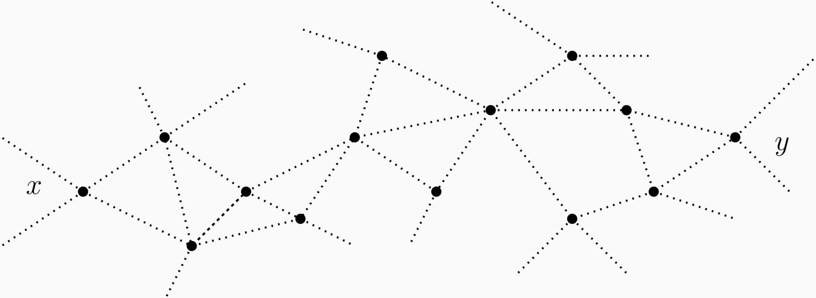

Figure 12: Enhancing the Isomap algorithm (Aamari, B. & Levard, 2022).

▷ The accuracy becomes $\varepsilon_n \approx n^{-1/d}$: that's better.

Metric learning

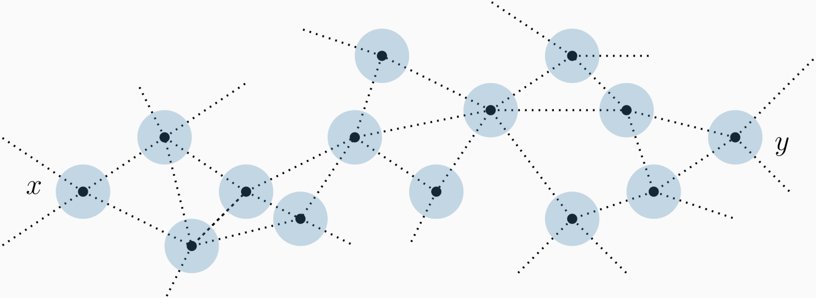

Figure 12: Enhancing the Isomap algorithm (Aamari, B. & Levard, 2022).

▷ The accuracy becomes $\varepsilon_n \approx n^{-1/d}$: that's better.

Metric learning

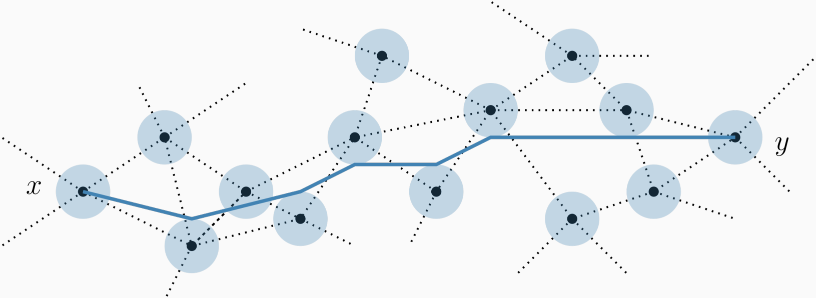

Figure 12: Enhancing the Isomap algorithm (Aamari, B. & Levard, 2022).

- The accuracy becomes $\varepsilon_n \approx n^{-1/d}$: that's better.

Metric learning

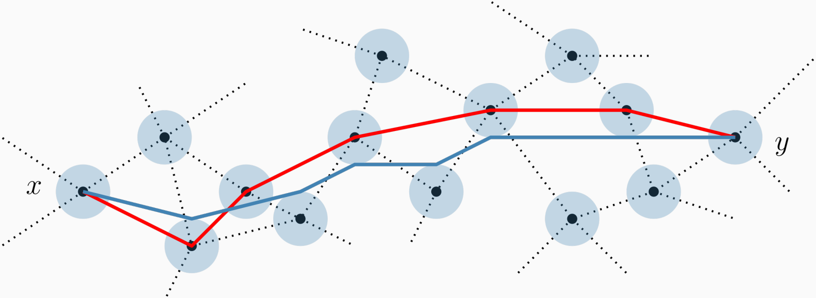

Figure 12: Enhancing the Isomap algorithm (Aamari, B. & Levard, 2022).

▷ The accuracy becomes $\varepsilon_n \approx n^{-1/d}$: that's better.

Metric learning

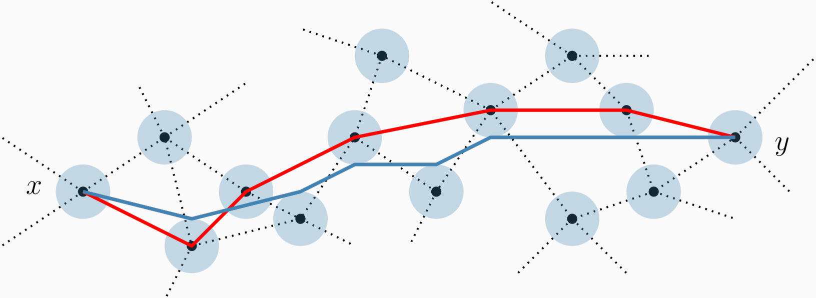

Figure 12: Enhancing the Isomap algorithm (Aamari, B. & Levard, 2022).

- ▷ The accuracy becomes $\varepsilon_n \approx n^{-1/d}$: that's better.

More generally, building upon polynomial patches instead of metric graphs, one can get:

Theorem Aamari, B. & Levrard (2022)

There exists an estimator $\widehat{\mathrm{d}}$ such that, for any $P \in \Sigma_k$, uniformly for any $x, y \in M$ where $M = \operatorname{supp} P$, there holds

$$(1-\varepsilon_n)\widehat{\mathrm{d}}(x,y) \leqslant \mathrm{d}_M(x,y) \leqslant (1+\varepsilon_n)\widehat{\mathrm{d}}(x,y)$$

with high probability and with $\varepsilon_n \approx n^{-k/d}$.

Furthermore, this accuracy is optimal.

Optimal reach estimation

1. The reach: definition and model
2. Estimation strategies for the reach
3. Optimal metric learning
4. Optimal reach estimation
5. Conclusion

Estimating the distortion radius

Idea: plug-in estimation of the sdr.

$$\widehat{\mathrm{sdr}} := \sup\left\{r \mid \forall x, y \in \widehat{M},\ \Delta \leqslant \|x-y\| \leqslant 2r \Rightarrow \widehat{\mathrm{d}}(x,y) \leqslant \mathrm{d}_{\mathscr{S}(r)}(x,y)\right\}.$$

- ▷ The sdr needs to be stable with respect to small perturbations of (M, d_M).

Two embedded spaces (K, d) and (K', d') are (ε, ν)-close if

1. $\mathrm{d}_{\mathrm{H}}(K, K') \leqslant \varepsilon$;
2. $\forall x, y \in K$ that are Δ-apart

$$(1-\nu)\mathrm{d}'(x', y') \leqslant \mathrm{d}(x,y) \leqslant (1+\nu)\mathrm{d}'(x', y')$$

for all x', y' among the nearest neighbors of x and y in K.

Estimating the distortion radius

Idea: plug-in estimation of the sdr.

$$\widehat{\mathrm{sdr}} := \sup\left\{r \mid \forall x, y \in \widehat{M},\ \Delta \leqslant \|x-y\| \leqslant 2r \Rightarrow \widehat{\mathrm{d}}(x,y) \leqslant \mathrm{d}_{\mathscr{S}(r)}(x,y)\right\}.$$

- The sdr needs to be stable with respect to small perturbations of (M, d_M).

Two embedded spaces (K, d) and (K', d') are (ε, ν)-close if

1. $\mathrm{d}_{\mathrm{H}}(K, K') \leqslant \varepsilon$;
2. $\forall x, y \in K$ that are Δ-apart

$$(1-\nu)\mathrm{d}'(x', y') \leqslant \mathrm{d}(x,y) \leqslant (1+\nu)\mathrm{d}'(x', y')$$

for all x', y' among the nearest neighbors of x and y in K.

Estimating the distortion radius

For any reasonable embedded metric space (K, d) (e.g. (M, d_M) where M is a submanifold):

Proposition

For any other space (K', d') that is (ε, ν)-close to (K, d), there holds

$$|\mathrm{sdr}_\Delta(K, \mathrm{d}) - \mathrm{sdr}_\Delta(K', \mathrm{d}')| \preccurlyeq \frac{\varepsilon \vee \Delta\nu}{\Delta^4}.$$

- We show that if $\mathrm{d_H}(M, \widehat{M}) \leqslant \varepsilon$ and $\widehat{\mathrm{d}}$ is the estimator described before, then $(\widehat{M}, \widehat{\mathrm{d}})$ is $(\varepsilon, \varepsilon/\Delta)$-close to (M, d_M).

Optimal reach estimation

Using local polynomial patching (Aamari and Levrard, 2019) yields:

Theorem Aamari, B. & Levrard (2022)

For any $k \geqslant 3$, there exists an estimator $\widehat{\mathrm{sdr}}$ such that

$$\sup_{P \in \Sigma_k} \mathbb{E}_{P^{\otimes n}} \left[|\widehat{\mathrm{sdr}} - \mathrm{sdr}_\Delta(M)| \right] \preccurlyeq \Delta^{-4} n^{-\frac{k}{d}},$$

and this rate is minimax-optimal.

- Faster rate than for the local reach !
- Diverging risk as $\Delta \to 0$ (which is expected since $\mathrm{sdr}_\Delta(M) \to \mathrm{rch}(M)$ then).

Optimal reach estimation

Letting Σ_k^α be the submodel on which

$$R_\ell(M) - \mathrm{wfs}(M) \geqslant \alpha,$$

there holds for the estimator

$$\widehat{\mathrm{rch}} = \widehat{R_\ell} \wedge \widehat{\mathrm{sdr}},$$

Theorem Aamari, B. & Levrard (2022)

For any $k \geqslant 3$, adaptively on $\alpha \in \mathbb{R}$,

$$\forall \alpha > 0, \quad \sup_{P \in \Sigma_k^\alpha} \mathbb{E}_{P^{\otimes n}} \left[|\widehat{\mathrm{rch}} - \mathrm{rch}(M)| \right] \preccurlyeq n^{-\frac{k}{d}},$$

$$\forall \alpha \leqslant 0, \quad \sup_{P \in \Sigma_k^\alpha} \mathbb{E}_{P^{\otimes n}} \left[|\widehat{\mathrm{rch}} - \mathrm{rch}(M)| \right] \preccurlyeq n^{-\frac{k-2}{d}},$$

and these rates are minimax-optimal.

Conclusion

Summary of the strategy

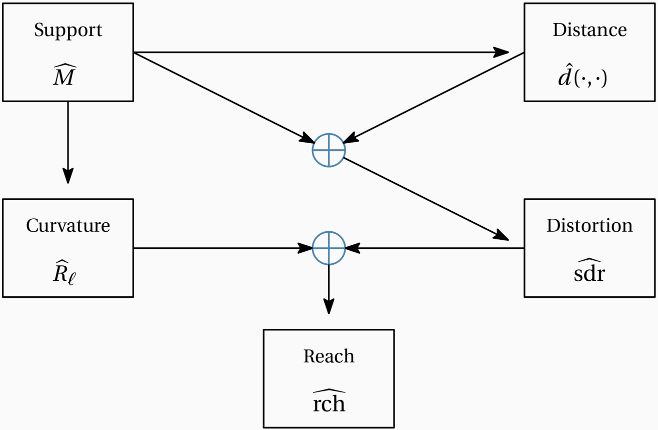

Figure 13: The optimal reach estimation pipeline

Summary of the strategy

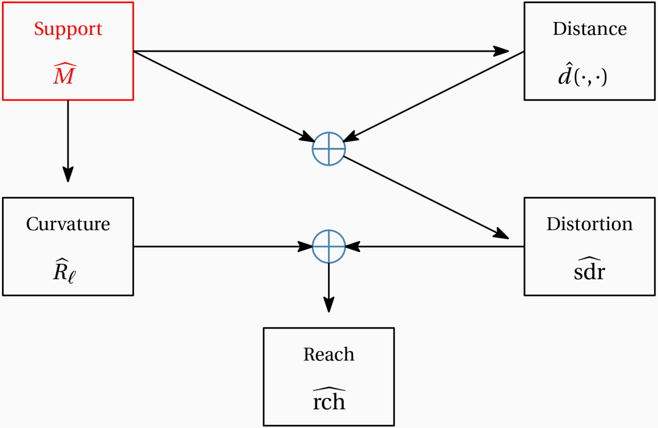

Figure 13: The optimal reach estimation pipeline

Summary of the strategy

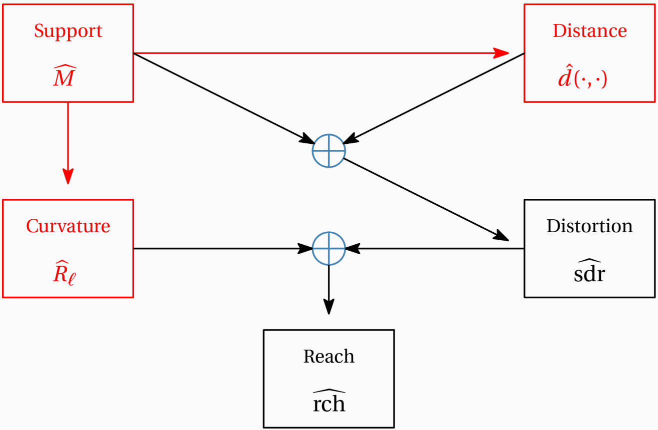

Figure 13: The optimal reach estimation pipeline

Summary of the strategy

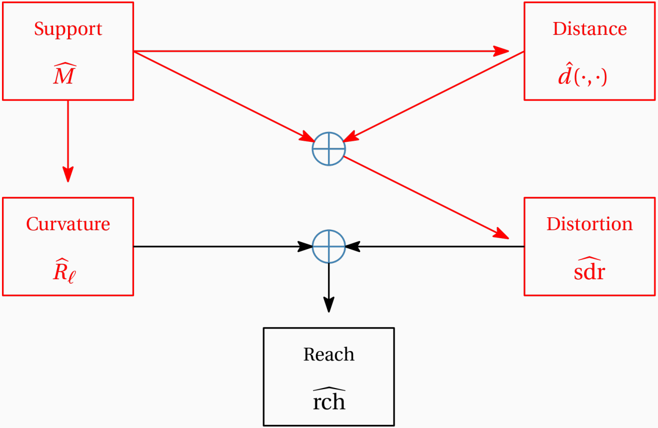

Figure 13: The optimal reach estimation pipeline

Summary of the strategy

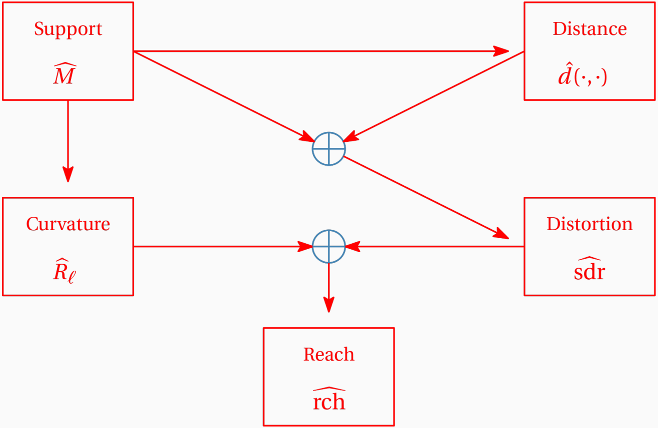

Figure 13: The **optimal** reach estimation pipeline

Conclusion and prospects

To sum up:

- ▷ Optimal estimation rates for the reach;
- ▷ Estimation of other geometric quantities along the way;
- ▷ Optimal estimation of geodesic lengths.

Possible developments:

- ▷ Computationally efficient estimation procedures;
- ▷ Geometric estimations under additive noise.

Thank you for your attention!

▹ Aamari, B. & Levrard (2022). Optimal Reach Estimation and Metric Learning. *arXiv preprint arXiv:2207.06074*.