

Statistical Theory for Manifold Learning and Generative Modeling

EDDIE AAMARI

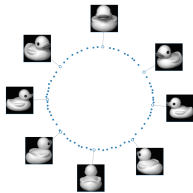
DÉPARTEMENT DE MATHÉMATIQUES ET APPLICATIONS
CNRS, ÉCOLE NORMALE SUPÉRIEURE PSL

HABILITATION À DIRIGER DES RECHERCHES

SEPTEMBER 21ST, 2026

1

Data geometry
and regularity



2

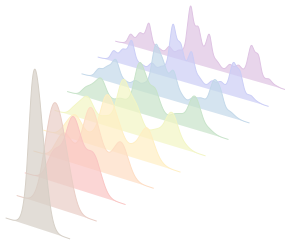
Manifold
learning



$$X_1, \dots, X_n \mapsto \hat{M}$$

3

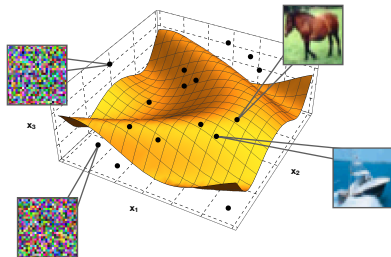
Generative
modeling



$$X_1, \dots, X_n \mapsto \hat{X}_{n+1}$$

THE MANIFOLD HYPOTHESIS

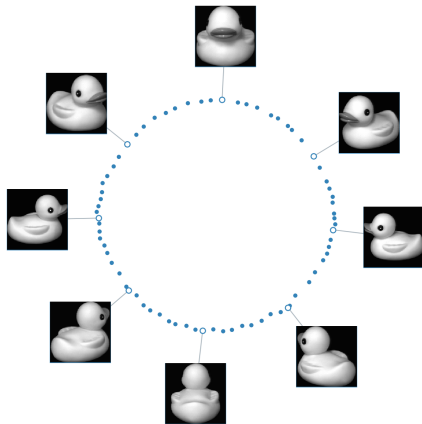
Manifold hypothesis: high-dimensional data lie close to low-dimensional structures.



[Goldt et al., 2020]

- Local nonlinear relations between the measured coordinates.
- A low-dimensional structure independent of the coordinate system.

A ROTATING OBJECT TRACES A CLOSED CURVE



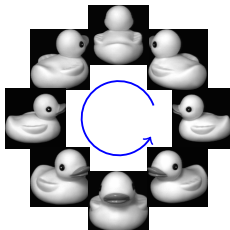
$D = 16\,384$ pixels, one viewing angle $\theta \in \mathbb{S}^1$.

[Nene–Nayar–Murase, 1996] COIL-20; Isomap from pixel distances.

MANIFOLD STRUCTURE IN IMAGES

Rotating objects

$D = 16\,384$ pixels



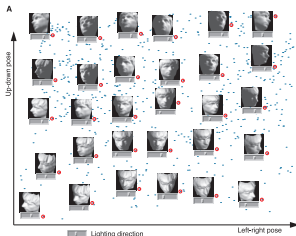
$d = 1$

Viewing angle

[Nene–Nayar–Murase, 1996]

Pose and light

$D = 4\,096$ pixels



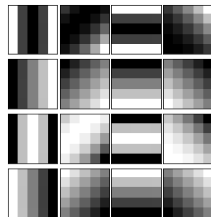
$d = 3$

Two pose angles + light

[Tenenbaum–de Silva–
Langford, 2000]

Natural image patches

3×3 pixels, normalized



$d = 2$

High-density patch model

[Carlsson et al., 2008]
Filters: [Love et al., 2023]

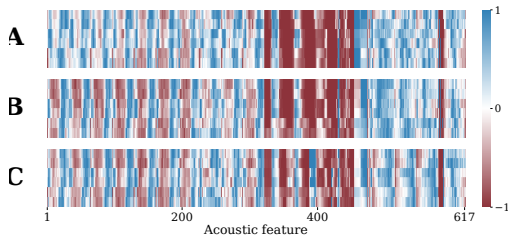
Many pixels, few degrees of freedom.

HIGHER-DIMENSIONAL DATA MANIFOLDS

Spoken letters

$D = 617$ acoustic features

One row = one spoken letter



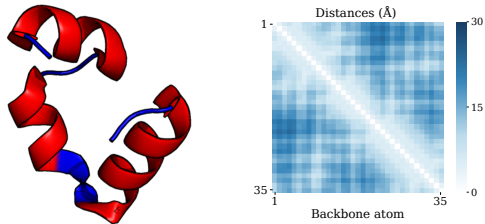
$$\hat{d}_{\text{MLE}} \simeq 17 \ll 617$$

Shared anatomy and articulation

ISOLET, Table 4: [Gomtsyan et al., 2019]

Protein conformations

$D = 595$ distances per conformation



$$\hat{d} \simeq 12-15 \ll 595$$

Folded villin, pairwise C α distances

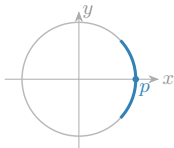
[Cazzaniga–Giorgino, 2026]

Many measurements, few local directions of variation.

WORTSCHATZ

A d -dimensional $\mathcal{C}^k$ submanifold $M \subset \mathbb{R}^D$ is locally the image of a $\mathcal{C}^k$ embedding of $\mathbb{R}^d$.

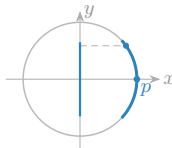
Implicit equations



$$x^2 + y^2 - 1 = 0$$

Correlations

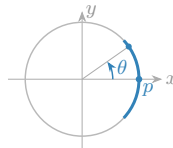
Local graphs



$$x = \sqrt{1 - y^2}$$

Regression

Local coordinates



$$\theta \mapsto (\cos \theta, \sin \theta)$$

Latent variables

FROM DATA EXAMPLES TO STATISTICAL QUESTIONS

An idealized observation model:

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} P, \quad P \in \mathcal{P}.$$

$$M := \text{supp } P$$

Model $\mathcal{P}$ encodes notions of regularity: dimension, smoothness, shape constraints, etc.

What can n observations determine under these assumptions?

Learn the geometry

$$\widehat{M} \simeq M.$$

Learn to generate

$$\widehat{P} \approx P.$$

WHY ASK FOR OPTIMAL RATES?

For a model $\mathcal{P}$, a target $\theta(P)$ and a loss ℓ , the minimax risk is

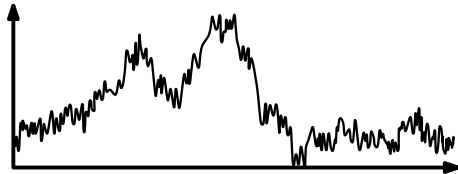
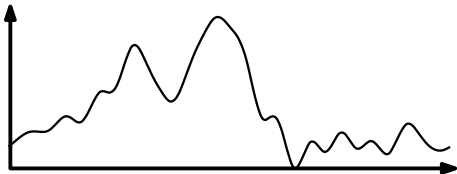
$$R_n(\mathcal{P}) = \inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{\otimes n}} \left[\ell(\theta(P), \hat{\theta}_n(X_1, \dots, X_n)) \right].$$

Upper bound. Which information suffices to construct an estimator?

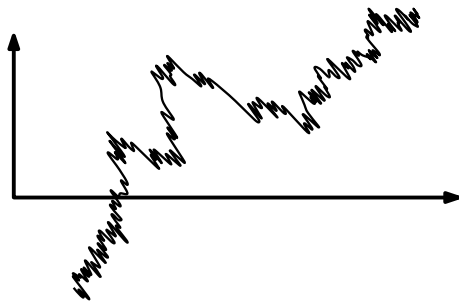
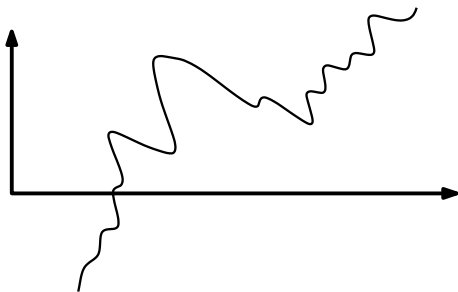
Lower bound. Which perturbations remain statistically indistinguishable?

Matching the bounds identifies which assumptions determine the achievable accuracy.

STATISTICS WITHOUT PREFERRED COORDINATES



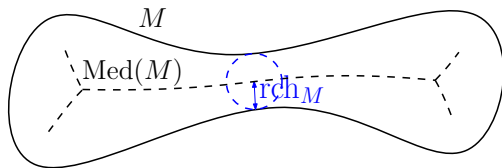
STATISTICS WITHOUT PREFERRED COORDINATES



Rotations preserve the geometry, but may destroy a graph over fixed coordinates.

Better not to choose a coordinate system?

REACH: A UNIFORM GEOMETRIC SCALE



[Federer, 1959] The **reach** of M is

$$\text{rch}_M = \text{dist}(M, \text{Med}(M)).$$

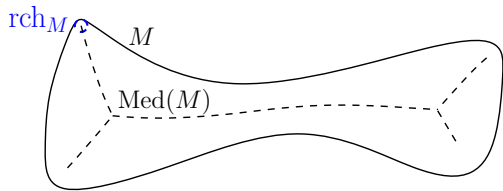
The medial axis consists of points with several nearest neighbors in M .

Within distance rch_M : unique projection.

For smooth manifolds: curvature and separation of branches.

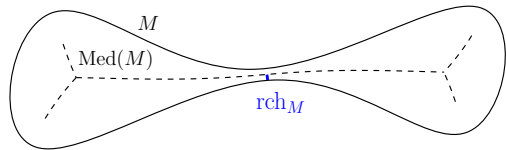
WHAT MAKES THE REACH SMALL?

Curvature



“ $M \in \mathcal{C}^2(1/\text{rch}_M)$...

Bottlenecks



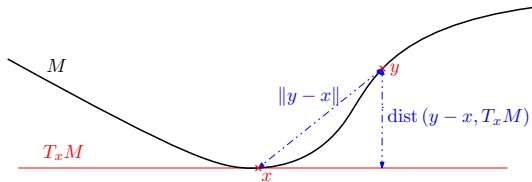
... Within ambient balls of size rch_M ”

The reach is attained by curvature or by a bottleneck.

[Aamari, Kim et al., 2019]

INSIGHTS AND ESTIMATION OF THE REACH

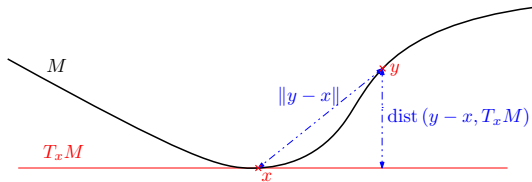
- [Federer, 1959] For all $x, y \in M$, $\text{dist}(y - x, T_x M) \leq \frac{\|y - x\|^2}{2\text{rch}_M}$.



INSIGHTS AND ESTIMATION OF THE REACH

- [Federer, 1959] For all $x, y \in M$, $\text{dist}(y - x, T_x M) \leq \frac{\|y - x\|^2}{2\text{rch}_M}$. Furthermore,

$$\frac{1}{\text{rch}_M} = \sup_{x \neq y \in M} \frac{2\text{dist}(y - x, T_x M)}{\|y - x\|^2}$$

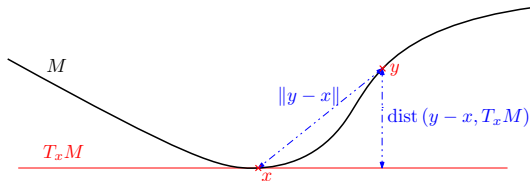


INSIGHTS AND ESTIMATION OF THE REACH

- [Federer, 1959] For all $x, y \in M$, $\text{dist}(y - x, T_x M) \leq \frac{\|y - x\|^2}{2\text{rch}_M}$. Furthermore,

$$\frac{1}{\text{rch}_M} = \sup_{x \neq y \in M} \frac{2\text{dist}(y - x, T_x M)}{\|y - x\|^2}$$

$\hookrightarrow$ Plugin estimator of rch_M using tangent spaces [Aamari, Kim et al., 2019]

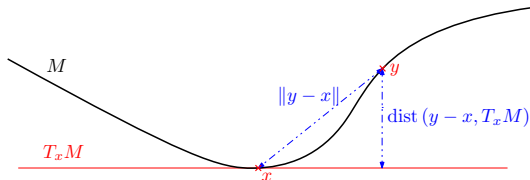


INSIGHTS AND ESTIMATION OF THE REACH

- [Federer, 1959] For all $x, y \in M$, $\text{dist}(y - x, T_x M) \leq \frac{\|y - x\|^2}{2\text{rch}_M}$. Furthermore,

$$\frac{1}{\text{rch}_M} = \sup_{x \neq y \in M} \frac{2\text{dist}(y - x, T_x M)}{\|y - x\|^2}$$

$\hookrightarrow$ Plugin estimator of rch_M using tangent spaces [Aamari, Kim et al., 2019]



- [Boissonnat–Lieutier–Wintraecken, 2019]

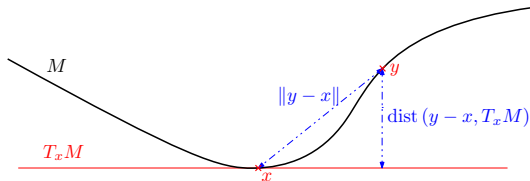
$$\text{rch}_M = \sup \left\{ r > 0 \mid \forall x, y \in M, \|x - y\| < 2r \Rightarrow d_M(x, y) \leq 2r \arcsin(\|y - x\|/2r) \right\}.$$

INSIGHTS AND ESTIMATION OF THE REACH

- [Federer, 1959] For all $x, y \in M$, $\text{dist}(y - x, T_x M) \leq \frac{\|y - x\|^2}{2\text{rch}_M}$. Furthermore,

$$\frac{1}{\text{rch}_M} = \sup_{x \neq y \in M} \frac{2\text{dist}(y - x, T_x M)}{\|y - x\|^2}$$

$\hookrightarrow$ Plugin estimator of rch_M using tangent spaces [Aamari, Kim et al., 2019]

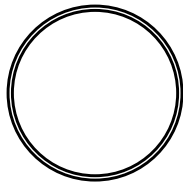


- [Boissonnat–Lieutier–Wintraecken, 2019]

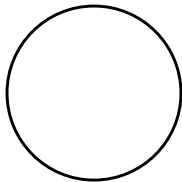
$$\text{rch}_M = \sup \left\{ r > 0 \mid \forall x, y \in M, \|x - y\| < 2r \Rightarrow d_M(x, y) \leq 2r \arcsin(\|y - x\|/2r) \right\}.$$

$\hookrightarrow$ Plugin estimator of rch_M using geodesic distances [Aamari–Berenfeld–Levrard, 2023]

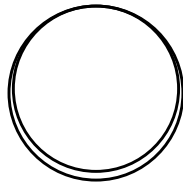
WHAT CAN GO WRONG WITH NEARLY COINCIDENT SHEETS?



$$\beta_0 = 2; \text{Vol} \simeq 2.$$



$$\beta_0 = 1; \text{Vol} = 1.$$



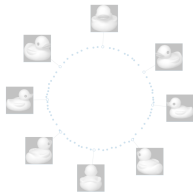
$$\beta_0 = 1; \text{Vol} \simeq 2.$$

Microscopic curvature and Hausdorff variations with jumps in topology and volume.

Hausdorff accuracy alone does not control every geometric quantity.

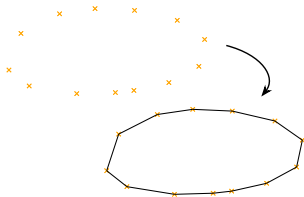
1

Data geometry
and regularity



2

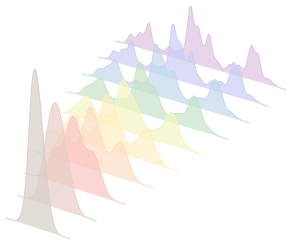
Manifold
learning



$$X_1, \dots, X_n \mapsto \hat{M}$$

3

Generative
modeling



$$X_1, \dots, X_n \mapsto \hat{X}_{n+1}$$

LEARNING THE SUPPORT OF A DISTRIBUTION

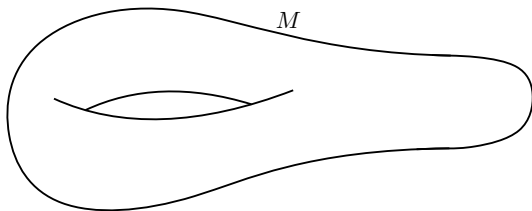
$$X_1, \dots, X_n \sim P \quad \longrightarrow \quad \widehat{M} \approx M := \text{supp} P.$$

LEARNING THE SUPPORT OF A DISTRIBUTION

$$X_1, \dots, X_n \sim P \quad \longrightarrow \quad \widehat{M} \approx M := \text{supp} P.$$

Assume first that P satisfies:

- **(smoothness)** M is a compact connected d -dimensional $\mathcal{C}^2$ submanifold without boundary,
- **(bounded reach)** $\text{rch}_M \geq \text{rch}_{\min} > 0$,
- **(bounded density)** $0 < f_{\min} \leq \text{d}P/\text{d}\mathcal{H}_M^d \leq f_{\max} < \infty$.



Denote by $\mathcal{P}_{\text{rch}}$ the set of distributions satisfying these conditions.

A RECONSTRUCTION THEOREM

Theorem [Aamari–Lévrard, 2019]

If $P \in \mathcal{P}_{\text{rch}}$, one can build an estimator $\hat{M}_{\text{PATCH}}$ based on data points $X_1, \dots, X_n$ such that

$$d_{\text{H}}(M, \hat{M}_{\text{PATCH}}) \leq C \left(\frac{\log n}{n} \right)^{2/d} \text{ w.h.p..}$$

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

A RECONSTRUCTION THEOREM

Theorem [Aamari–Levrard, 2019]

If $P \in \mathcal{P}_{\text{rch}}$, one can build an estimator $\hat{M}_{\text{PATCH}}$ based on data points $X_1, \dots, X_n$ such that

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

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

→ Other estimators achieving the same Hausdorff rate:

- Empirical risk manifold minimizer [Genovese et al., 2012]
- Local Tangent Delaunay triangulation [Aamari–Levrard, 2018]
- Local convex hulls [Divol, 2021]

INGREDIENT I: APPROXIMATION THEORY

Theorem [Aamari–Levrard, 2019]

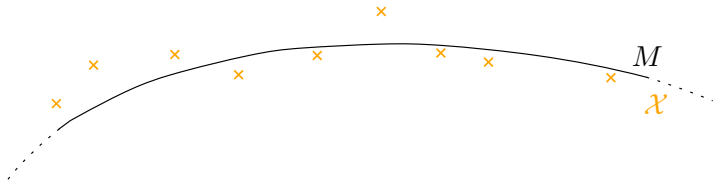
For $\Delta \lesssim \text{rch}_{\min}$, assume that we have a point cloud $\mathcal{X} \subset \mathbb{R}^D$ that is:

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

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

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

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



INGREDIENT I: APPROXIMATION THEORY

Theorem [Aamari–Levrard, 2019]

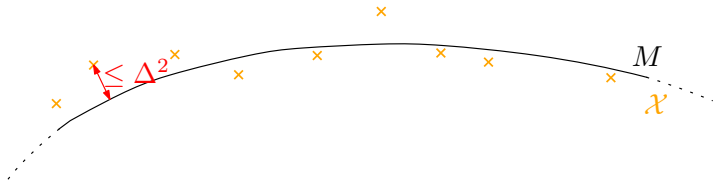
For $\Delta \lesssim \text{rch}_{\min}$, assume that we have a point cloud $\mathcal{X} \subset \mathbb{R}^D$ that is:

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

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

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

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



INGREDIENT I: APPROXIMATION THEORY

Theorem [Aamari–Levrard, 2019]

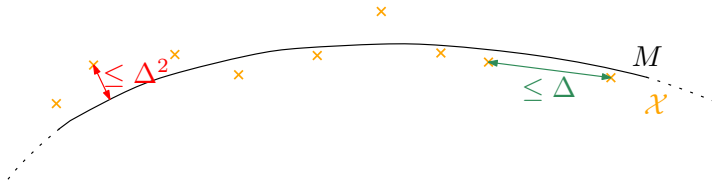
For $\Delta \lesssim \text{rch}_{\min}$, assume that we have a point cloud $\mathcal{X} \subset \mathbb{R}^D$ that is:

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

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

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

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



INGREDIENT I: APPROXIMATION THEORY

Theorem [Aamari–Levrard, 2019]

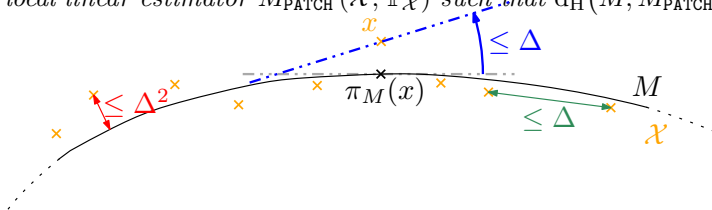
For $\Delta \lesssim \text{rch}_{\min}$, assume that we have a point cloud $\mathcal{X} \subset \mathbb{R}^D$ that is:

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

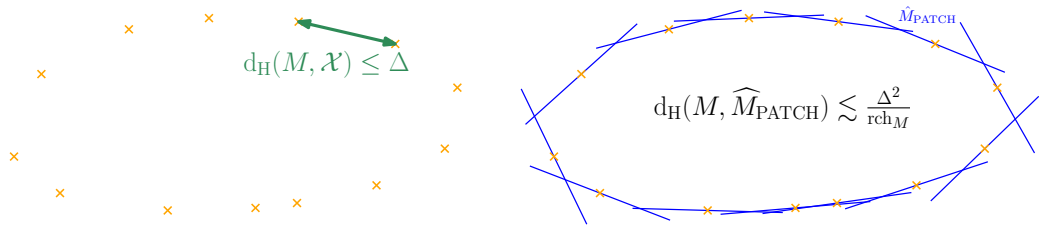
together with a family $\mathbb{T}_{\mathcal{X}}$ of linear spaces that

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

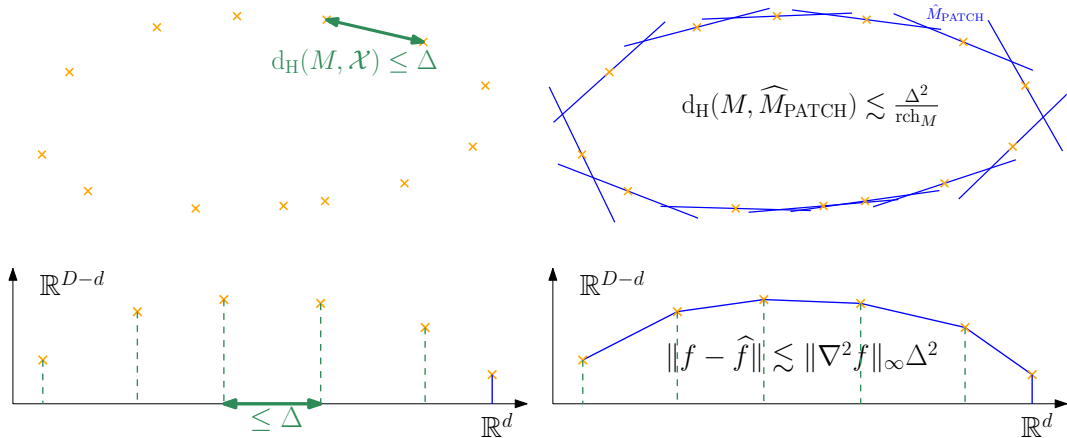
One can build a local linear estimator $\hat{M}_{\text{PATCH}}(\mathcal{X}, \mathbb{T}_{\mathcal{X}})$ such that $d_H(M, \hat{M}_{\text{PATCH}}) \lesssim \Delta^2 / \text{rch}_{\min}.$



INGREDIENT I: APPROXIMATION THEORY



INGREDIENT I: APPROXIMATION THEORY



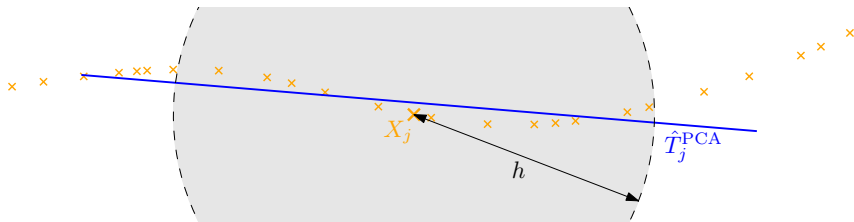
INGREDIENT II: LOCAL PCA

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

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

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

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

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–Lévrard, 2019]

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

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

MANIFOLD ESTIMATION FROM A RANDOM SAMPLE

Proposition [Aamari–Levrard, 2019]

With high probability, an i.i.d. sample $\mathbb{X}_n = \{X_1, \dots, X_n\}$ from $P \in \mathcal{P}_{\text{rch}}$ fulfills:

$$\begin{array}{ll} - \max_{X_j \in \mathbb{X}_n} \text{dist}(X_j, M) = 0 & - \sup_{p \in M} \text{dist}(p, \mathbb{X}_n) \lesssim (\log n/n)^{1/d}. \end{array}$$

The family of d -planes $(\hat{T}_j^{\text{PCA}})_{j=1}^n$ built from local PCA fulfills

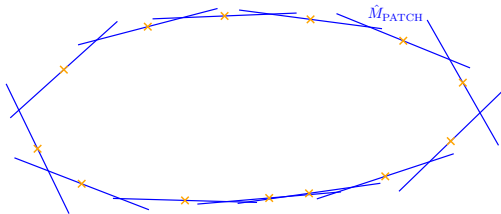
$$\max_{X_j \in \mathbb{X}_n} \angle(T_{X_j} M, \hat{T}_j^{\text{PCA}}) \lesssim (\log n/n)^{1/d}.$$

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

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

This rate is minimax optimal.

MANIFOLD ESTIMATION UNDER REACH CONDITION

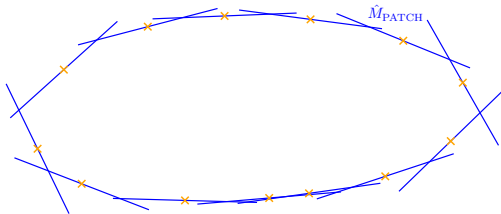


Theorem [Aamari–Levrard, 2019]

Under the reach condition, local PCA yields a union of affine patches such that

$$\sup_{P \in \mathcal{P}_{\text{rch}}} \mathbb{E} \, d_{\text{H}}(M, \widehat{M}_{\text{PATCH}}) \lesssim \left(\frac{\log n}{n} \right)^{2/d}.$$

MANIFOLD ESTIMATION UNDER REACH CONDITION



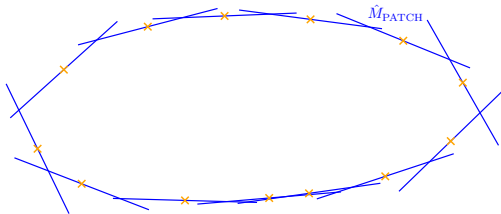
Theorem [Aamari–Levrard, 2019]

Under the reach condition, local PCA yields a union of affine patches such that

$$\sup_{P \in \mathcal{P}_{\text{rch}}} \mathbb{E} d_H(M, \widehat{M}_{\text{PATCH}}) \lesssim \left(\frac{\log n}{n} \right)^{2/d}.$$

- ↪ Possible to recover the exact topology of M with triangulations [Aamari–Levrard, 2018]
- ↪ Possible to adapt to $\mathcal{C}^k$ smoothness with rate $(\log n/n)^{k/d}$ [Aamari–Levrard, 2019]
- ↪ Possible to adapt to $\partial M \neq \emptyset$, with rate $(\log n/n)^{2/(d+1)}$ [Aamari–Aaron–Levrard, 2023]
- ↪ Possible to impose differential privacy and handle adversarial noise [Aamari–Knop, 2024]

MANIFOLD ESTIMATION UNDER REACH CONDITION



Theorem [Aamari–Levrard, 2019]

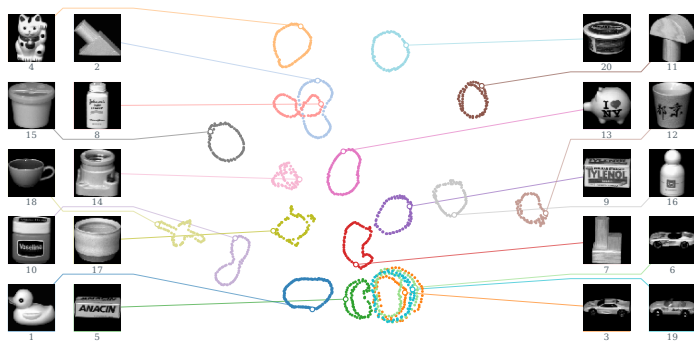
Under the reach condition, local PCA yields a union of affine patches such that

$$\sup_{P \in \mathcal{P}_{\text{rch}}} \mathbb{E} d_H(M, \widehat{M}_{\text{PATCH}}) \lesssim \left(\frac{\log n}{n} \right)^{2/d}.$$

- ↪ Possible to recover the exact topology of M with triangulations [Aamari–Levrard, 2018]
- ↪ Possible to adapt to $\mathcal{C}^k$ smoothness with rate $(\log n/n)^{k/d}$ [Aamari–Levrard, 2019]
- ↪ Possible to adapt to $\partial M \neq \emptyset$, with rate $(\log n/n)^{2/(d+1)}$ [Aamari–Aaron–Levrard, 2023]
- ↪ Possible to impose differential privacy and handle adversarial noise [Aamari–Knop, 2024]
- ↪ Possible to bypass the reach assumption! [Aamari–Berenfeld, 2024]

SEVERAL OBJECTS, SEVERAL VIEW TRAJECTORIES

20 objects, 72 views each: 1 440 images in $\mathbb{R}^{16384}$.



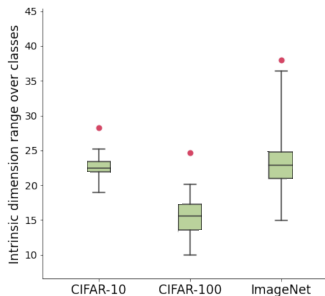
The three toy cars trace nearly overlapping loops in this UMAP view.

[Nene–Nayar–Murase, 1996] COIL-20; cosine pixel distance.

BEYOND THE MANIFOLD HYPOTHESIS

Varying dimensions

Dimension estimates across image classes



Boxplots: classes red dots: whole dataset

[Brown et al., 2023]

Self-intersections

Must different classes stay apart?



Chihuahua or muffin?

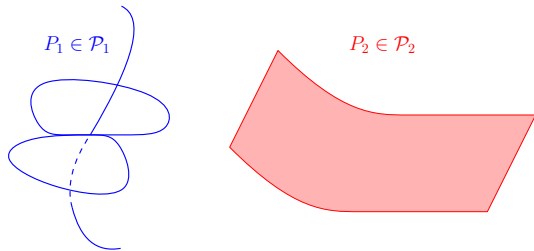
[Know Your Meme, 2016]

Need to allow different dimensions and branches that meet.

STATISTICAL MODEL WITHOUT REACH CONDITION

Single-source model: $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} P_d$, where $M_d = \text{supp}(P_d) \subset \mathbb{R}^D$ satisfies:

- (smoothness) M_d is a compact immersed $\mathcal{C}^2$ manifold of dimension d , without boundary;
- (bounded curvature & volume) $\kappa(M_d) \leq \kappa_{\max}$ and $\mathbb{V}(M_d) \leq C_d$;
- (bounded density) $0 < f_{\min,d} \leq dP_d/d\mathcal{H}_{M_d}^d \leq f_{\max,d} < \infty$.

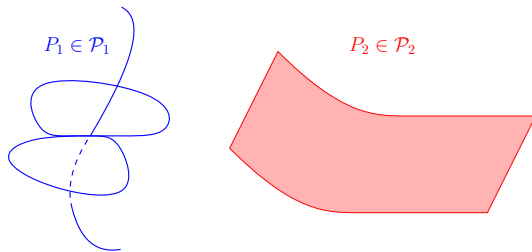


Denote by $\mathcal{P}_d$ the set of distributions satisfying these conditions.

STATISTICAL MODEL WITHOUT REACH CONDITION

Single-source model: $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} P_d$, where $M_d = \text{supp}(P_d) \subset \mathbb{R}^D$ satisfies:

- (smoothness) M_d is a compact immersed $\mathcal{C}^2$ manifold of dimension d , without boundary;
- (bounded curvature & volume) $\kappa(M_d) \leq \kappa_{\max}$ and $\mathbb{V}(M_d) \leq C_d$;
- (bounded density) $0 < f_{\min,d} \leq dP_d/d\mathcal{H}_{M_d}^d \leq f_{\max,d} < \infty$.

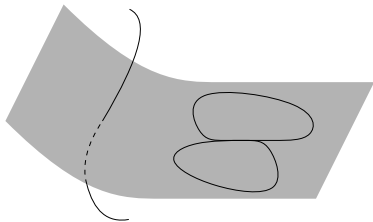


Denote by $\mathcal{P}_d$ the set of distributions satisfying these conditions.

At intersections, several directions can coexist

$\Rightarrow$ Local PCA fails.

UNION OF MANIFOLDS HYPOTHESIS



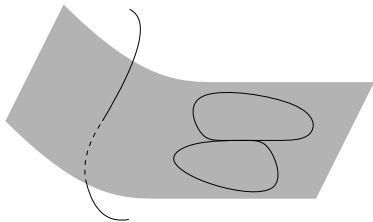
Mixture model: $P = \sum_{k=1}^K \alpha_k P_{d_k}, \quad P_{d_k} \in \mathcal{P}_{d_k}.$

$$1 \leq d_1 < \dots < d_K < D,$$

$$\alpha_k \geq \alpha_{\min} > 0, \quad \sum_k \alpha_k = 1.$$

Denote by $\mathcal{P}_{\text{mix}}$ the set of distributions satisfying these conditions.

UNION OF MANIFOLDS HYPOTHESIS

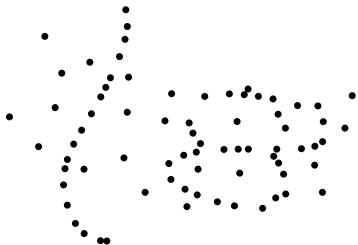


Mixture model: $P = \sum_{k=1}^K \alpha_k P_{d_k}, \quad P_{d_k} \in \mathcal{P}_{d_k}.$

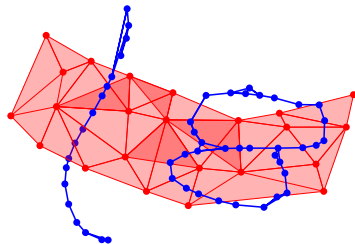
$$1 \leq d_1 < \dots < d_K < D,$$

$$\alpha_k \geq \alpha_{\min} > 0, \quad \sum_k \alpha_k = 1.$$

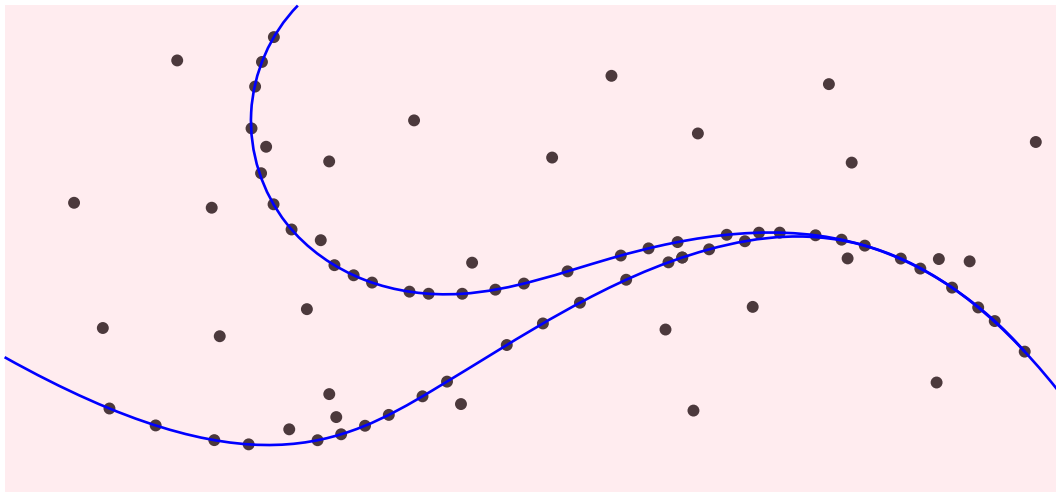
Denote by $\mathcal{P}_{\text{mix}}$ the set of distributions satisfying these conditions.



Stratification learning



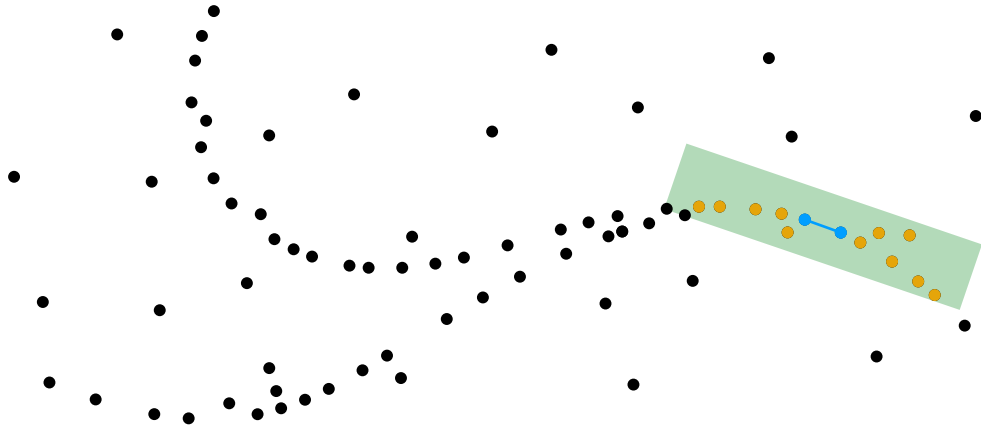
THE SLABELING ALGORITHM



THE SLABELING ALGORITHM



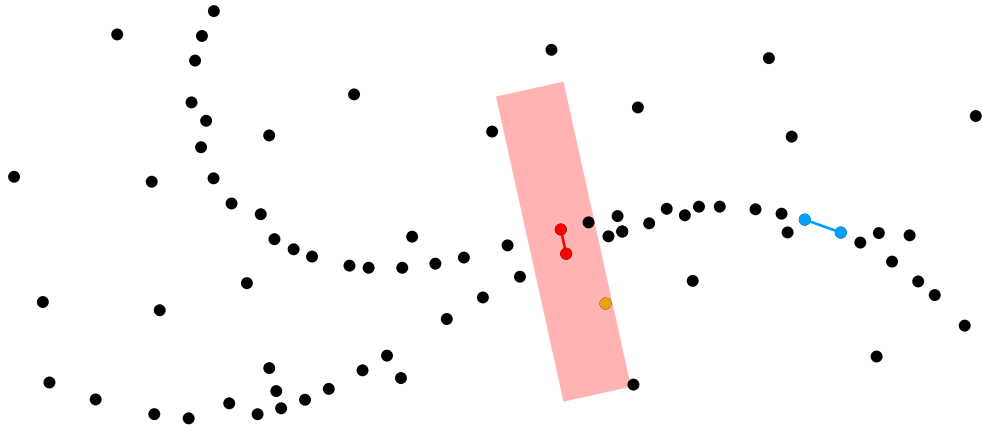
THE SLABELING ALGORITHM



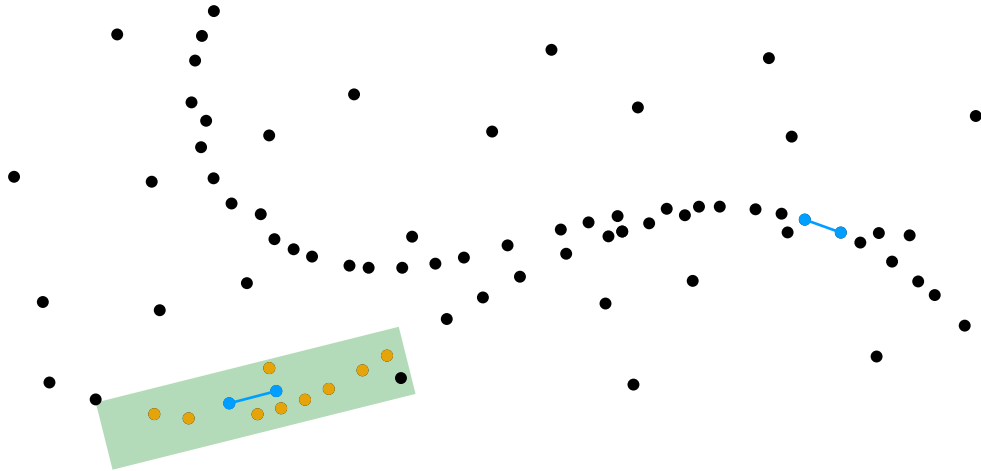
THE SLABELING ALGORITHM



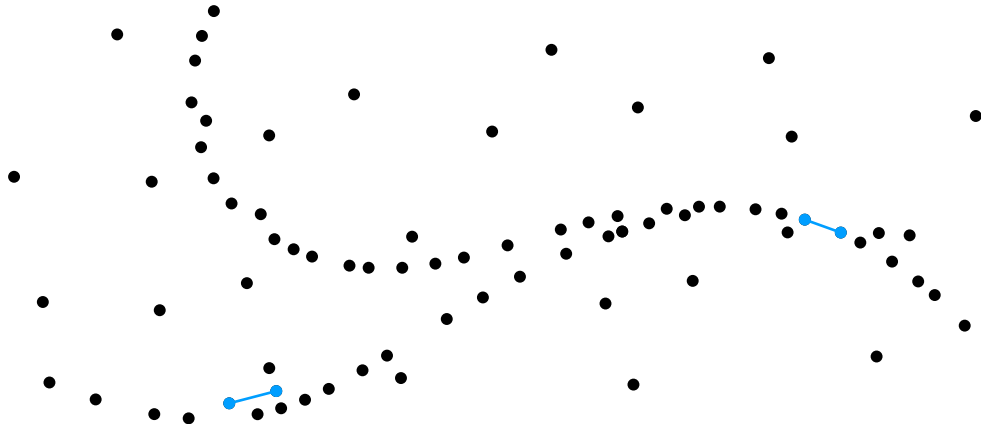
THE SLABELING ALGORITHM



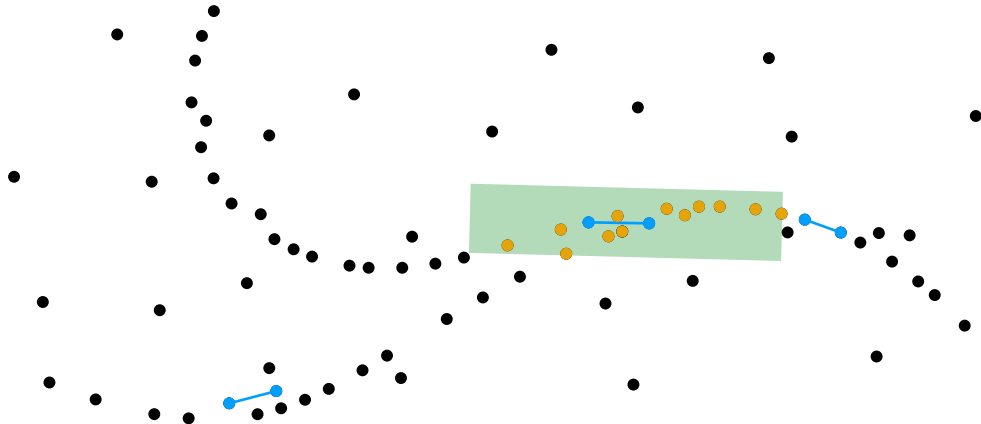
THE SLABELING ALGORITHM



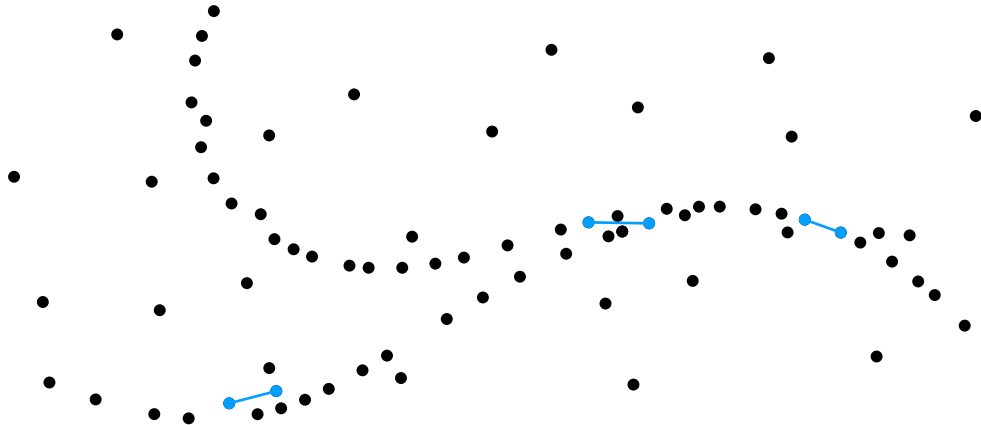
THE SLABELING ALGORITHM



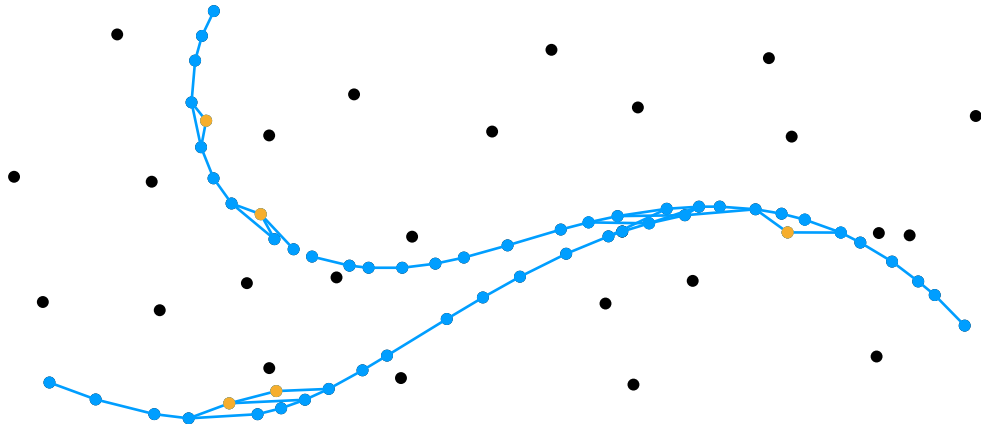
THE SLABELING ALGORITHM



THE SLABELING ALGORITHM



THE SLABELING ALGORITHM

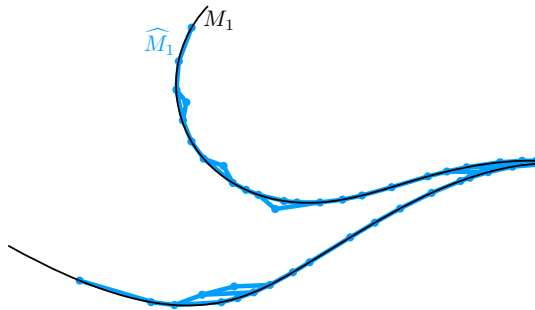


MANIFOLD ESTIMATION WITHOUT REACH CONDITION

Theorem [Aamari–Berenfeld, 2024]

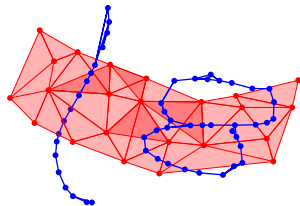
W.h.p., for all $k \in \{1, \dots, K\}$,

$$d_H(\widehat{M}_{d_k}, M_{d_k}) \lesssim \left(\frac{\log n}{n} \right)^{2/d_k}.$$



Also recovering:

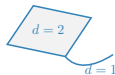
- Number of components K
- Active dimensions $d_1, \dots, d_K$
- Dimensionwise clustering
- Tangent cones up to displacement



A FEW KEY QUESTIONS IN MANIFOLD LEARNING

Stable models

Tractable assumptions preserved by mixtures and postprocessing?

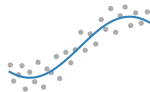


Noise

Which noise levels still allow geometric recovery?

[Genovese et al., 2012] [Puchkin–Spokoiny, 2022]

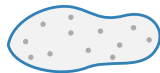
[Aizenbud–Sober, 2021] [Fefferman et al., 2023]



Volume estimation and uniform resampling

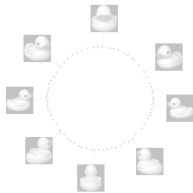
Can we do better than reconstructing the support?

[Baldin–Reiß, 2016] [Arias-Castro et al., 2019] [Brunel, 2014]



1

Data geometry
and regularity



2

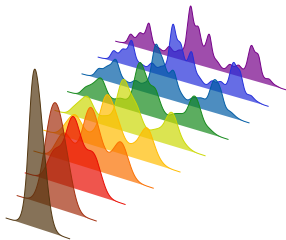
Manifold
learning



$$X_1, \dots, X_n \mapsto \hat{M}$$

3

Generative
modeling



$$X_1, \dots, X_n \mapsto \hat{X}_{n+1}$$

LEARNING TO RESAMPLE A DISTRIBUTION

$$X^{(1)}, \dots, X^{(n)} \sim P \quad \longrightarrow \quad \widehat{X}^{(n+1)} \approx P.$$

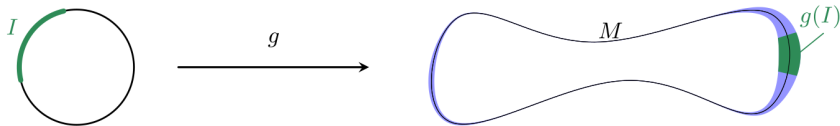
LEARNING TO RESAMPLE A DISTRIBUTION

$$X^{(1)}, \dots, X^{(n)} \sim P \quad \longrightarrow \quad \hat{X}^{(n+1)} \approx P.$$

GANs (used to...) search directly a pushforward map $g : \mathbb{R}^d \rightarrow \mathbb{R}^D$:

$$\hat{X}^{(n+1)} := \hat{g}(U^{(n+1)})$$

with $\hat{g}$ learnt adversarially.



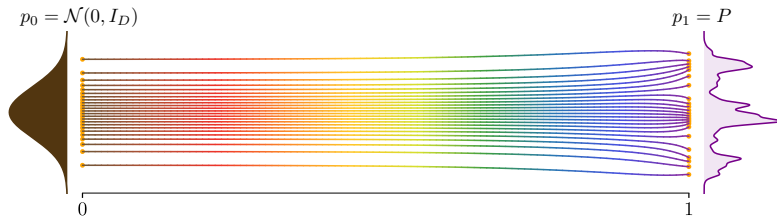
$\hookrightarrow$ **Minimax-optimal estimation in intrinsic dimension d**

[Stéphanovitch–Aamari–Levrard, 2024]

GENERATING BY A FLOW

Instead of fitting one map, **Markovian methods** learn the dynamics of moving particles:

$$\hat{X}^{(n+1)} := \hat{X}_{t=1} \text{ where } \begin{cases} \dot{\hat{X}}_t = \hat{v}_t(\hat{X}_t) \\ \hat{X}_{t=0} = G \sim \mathcal{N}(0, I_D) \end{cases}$$

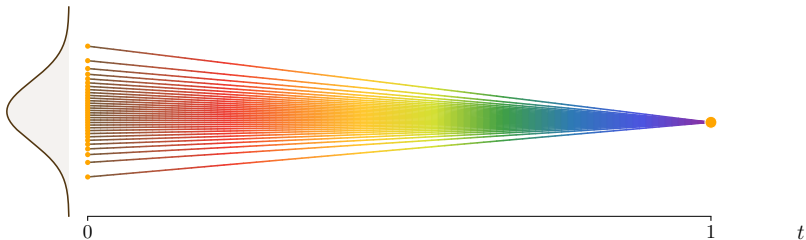


How can we learn a nicely behaved velocity?

GENERATOR MATCHING

Fix a target z , and draw $G \sim \mathcal{N}(0, \text{Id})$.

$$X_t^z = (1 - t)G + tz, \quad v_t^z(x) = \frac{z - x}{1 - t}, \quad t < 1.$$

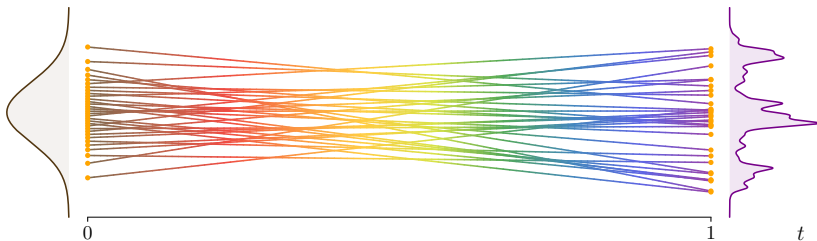


Every path goes straight to the same target z .

GENERATOR MATCHING

Draw $Z \sim P$ independently of $G \sim \mathcal{N}(0, \text{Id})$.

$$Y_t := X_t^Z = (1 - t)G + tZ, \quad p_t = \mathcal{L}(Y_t).$$

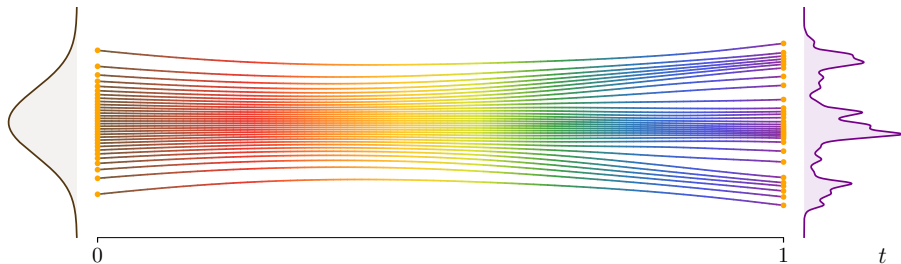


Every path keeps its own target Z .

GENERATOR MATCHING

For $t < 1$, follow $\dot{X}_t = v_t(X_t)$, with $X_0 \sim \mathcal{N}(0, \text{Id})$, where

$$v_t(x) = \mathbb{E}[v_t^Z(X_t^Z) \mid X_t^Z = x] = \mathbb{E}[Z - G \mid Y_t = x].$$



$$\mathcal{L}(X_t) = \mathcal{L}(Y_t) = p_t \text{ for every } t.$$

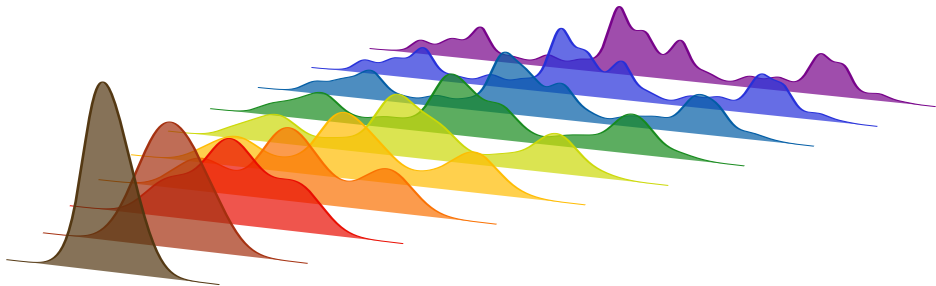
Same distributions at each time, different trajectories.

[Lipman et al., 2023]

GENERATOR MATCHING

For $t < 1$, follow $\dot{X}_t = v_t(X_t)$, with $X_0 \sim \mathcal{N}(0, \text{Id})$, where

$$v_t(x) = \mathbb{E}[v_t^Z(X_t^Z) \mid X_t^Z = x] = \mathbb{E}[Z - G \mid Y_t = x].$$



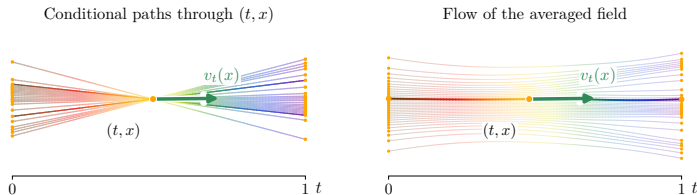
$$\mathcal{L}(X_t) = \mathcal{L}(Y_t) = p_t \text{ for every } t.$$

Same distributions at each time, different trajectories.

SCORE TRICK

The target velocity field has the conditional expectation form

$$v_t(x) = \mathbb{E}[v_t^Z(X_t^Z) \mid X_t^Z = x] .$$



$\Rightarrow$ On any interval I with finite square loss, for any candidate class $\mathcal{V}_I$,

$$\operatorname{argmin}_{w \in \mathcal{V}_I} \int_I \mathbb{E}_{X \sim p_t} [\|w_t(X) - v_t(X)\|^2] dt = \operatorname{argmin}_{w \in \mathcal{V}_I} \int_I \mathbb{E}_{Z, G} [\|w_t(X_t^Z) - v_t^Z(X_t^Z)\|^2] dt.$$

INGREDIENT I: VELOCITY MATCHING AND CONCENTRATION

Given $Z_1, \dots, Z_n \stackrel{\text{iid}}{\sim} P$, estimate on $I = [a, b] \subset (0, 1)$:

$$\hat{v}_I \in \operatorname{argmin}_{w \in \mathcal{V}_I} \frac{1}{n} \sum_{i=1}^n \int_I \mathbb{E}_{G \sim \mathcal{N}(0, I_D)} \left[\|w_t(X_t^{Z_i}) - v_t^{Z_i}(X_t^{Z_i})\|^2 \right] dt.$$

Proposition [\[Stéphanovitch–Aamari–Levrard, 2025\]](#)

For Gaussian paths, under suitable envelope and tail assumptions,

$$\int_I \|\hat{v}_{I,t} - v_t\|_{L^2(p_t)}^2 dt \lesssim \inf_{w \in \mathcal{V}_I} \int_I \|w_t - v_t\|_{L^2(p_t)}^2 dt + \frac{1}{n} \log[\mathcal{N}(\mathcal{V}_I, \|\cdot\|_\infty, n^{-1}) \vee n].$$

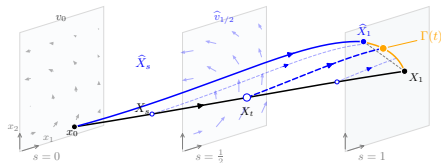
INGREDIENT II: STABILITY OF DYNAMICS

Proposition

Write $p_t = \mathcal{L}(X_t)$ and $\hat{p}_t = \mathcal{L}(\hat{X}_t)$, with $\dot{X}_t = v_t(X_t)$, $\dot{\hat{X}}_t = \hat{v}_t(\hat{X}_t)$, and $X_0 \sim \hat{X}_0$. Then,

$$W_1(p_1, \hat{p}_1) \leq \int_0^1 e^{\int_t^1 \hat{L}_s \, ds} \|v_t - \hat{v}_t\|_{L^1(p_t)} \, dt, \quad \hat{L}_s := \sup_x \lambda_{\max}(D\hat{v}_s(x)).$$

Here $\lambda_{\max}(A) := \sup_{\|u\|=1} \langle Au, u \rangle$.



Constrain the learned field so that $\int_0^1 (\hat{L}_s)_+ \, ds \lesssim 1$.

INGREDIENT III: INDUCED REGULARITY OF THE SCORE

The target distribution enters through posterior averaging:

$$p_t(x) = \int p_t^z(x) P(dz), \quad v_t(x) = \frac{\int v_t^z(x) p_t^z(x) P(dz)}{p_t(x)}.$$

For the following bounds, take $p_t^z = \mathcal{N}(tz, (1 - t^2)I_D)$, $0 < t < 1$:

$$v_t^z(x) = \frac{z - tx}{1 - t^2}, \quad v_t(x) = \frac{x + \nabla \log p_t(x)}{t}.$$

Proposition [\[Stéphanovitch, 2025\]](#)

Assume $p^\star = e^{-U+a}$, with U smooth and strongly convex, a is β -Hölder ($\beta > 0$), with controlled tails and derivatives.

$$\sup_x \lambda_{\max}(Dv_t(x)) \lesssim \frac{t}{(1 - t^2)^{1-\delta_\beta}}, \quad \delta_\beta := \frac{\beta \wedge 1}{2} > 0.$$

Moreover,

$$\|\partial_t^k D_x^j v_t\|_{L^\infty(A_t^\varepsilon)} \lesssim (1 - t^2)^{-\frac{(1+j+2k-\beta)_+}{2}}, \quad p_t(A_t^\varepsilon) \geq 1 - \varepsilon.$$

Smoother target $p^\star \Rightarrow$ Smoother $v_t \Rightarrow$ Faster estimation rates.

A STATISTICAL GUARANTEE FOR SCORE MODELS

Let $\mathcal{P}^\beta$ denote the preceding full-dimensional density model, with shape, tail and derivative bounds controlled.

Theorem [Stéphanovitch–Aamari–Levrard, 2025]

Score matching over suitable constrained neural classes, with tuned initialization and early stopping, gives minimax optimal rate

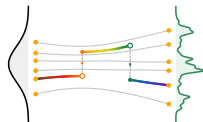
$$\sup_{p^\star \in \mathcal{P}^\beta} \mathbb{E} W_1(P, \hat{P}) \lesssim n^{-\frac{\beta+1}{2\beta+D}}, \quad P(\mathrm{d}x) = p^\star(x) \mathrm{d}x.$$

Same analysis for ODE sampling and SDE sampling.

A FEW KEY QUESTIONS IN GENERATIVE MODELING

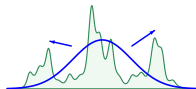
Generator matching

Beyond flows and diffusions: learn jump dynamics?



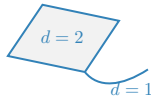
Drifting

Dynamical training, single-shot generation. Statistical accuracy?



Intrinsic data geometry

Sharp rates under weaker geometric assumptions?



WITH GRATITUDE TO ALL MY COAUTHORS

Catherine Aaron	Frédéric Chazal	Clément Levrard	Alessandro Rinaldo
Ery Arias-Castro	George Deligiannidis	Qinyu Li	Judith Rousseau
Iskander Azangulov	Jisu Kim	Bertrand Michel	Arthur Stéphanovitch
Clément Berenfeld	Alexander Knop	Peter Potapchik	Larry Wasserman

WITH GRATITUDE TO ALL MY COAUTHORS

Catherine Aaron	Frédéric Chazal	Clément Levrard	Alessandro Rinaldo
Ery Arias-Castro	George Deligiannidis	Qinyu Li	Judith Rousseau
Iskander Azangulov	Jisu Kim	Bertrand Michel	Arthur Stéphanovitch
Clément Berenfeld	Alexander Knop	Peter Potapchik	Larry Wasserman



TO BE
CONTINUED.