Covers and nerves: union of balls, geometric inference and Mapper

Frédéric Chazal
INRIA Saclay - Ile-de-France
frederic.chazal@inria.fr
Highlighting and inferring the topological structure of data

Idea:
Highlighting and inferring the topological structure of data

Idea:

• Group data points in “local clusters”
Highlighting and inferring the topological structure of data

Idea:

- Group data points in “local clusters”
- Summarize the data through the combinatorial/topological structure of intersection patterns of “clusters”
Highlighting and inferring the topological structure of data

Idea:
- Group data points in “local clusters”
- Summarize the data through the combinatorial/topological structure of intersection patterns of “clusters”

Goal: Do it in a way that preserves (some of) the topological features of the data.
Background mathematical notions

Topological space

A **topology** on a set $X$ is a family $\mathcal{O}$ of subsets of $X$ that satisfies the three following conditions:

1) the empty set $\emptyset$ and $X$ are elements of $\mathcal{O}$,
2) any union of elements of $\mathcal{O}$ is an element of $\mathcal{O}$,
3) any finite intersection of elements of $\mathcal{O}$ is an element of $\mathcal{O}$.

The set $X$ together with the family $\mathcal{O}$, whose elements are called open sets, is a **topological space**. A subset $C$ of $X$ is **closed** if its complement is an open set.

A map $f : X \rightarrow X'$ between two topological spaces $X$ and $X'$ is **continuous** if and only if the pre-image $f^{-1}(O') = \{x \in X : f(x) \in O'\}$ of any open set $O' \subset X'$ is an open set of $X$. Equivalently, $f$ is continuous if and only if the pre-image of any closed set in $X'$ is a closed set in $X$ (exercise).

A topological space $X$ is a **compact space** if any open cover of $X$ admits a finite subcover, i.e. for any family $\{U_i\}_{i \in I}$ of open sets such that $X = \bigcup_{i \in I} U_i$ there exists a finite subset $J \subseteq I$ of the index set $I$ such that $X = \bigcup_{j \in J} U_j$. 
Background mathematical notions

Metric space

A metric (or distance) on $X$ is a map $d : X \times X \to [0, +\infty)$ such that:

i) for any $x, y \in X$, $d(x, y) = d(y, x)$,

ii) for any $x, y \in X$, $d(x, y) = 0$ if and only if $x = y$,

iii) for any $x, y, z \in X$, $d(x, z) \leq d(x, y) + d(y, z)$.

The set $X$ together with $d$ is a metric space.

The smallest topology containing all the open balls $B(x, r) = \{y \in X : d(x, y) < r\}$ is called the metric topology on $X$ induced by $d$.

Example: the standard topology in an Euclidean space is the one induced by the metric defined by the norm: $d(x, y) = \|x - y\|$.

Compacity: a metric space $X$ is compact if and only if any sequence in $X$ has a convergent subsequence. In the Euclidean case, a subset $K \subset \mathbb{R}^d$ (endowed with the topology induced from the Euclidean one) is compact if and only if it is closed and bounded (Heine-Borel theorem).
Comparing topological spaces

Homeomorphy and isotopy

- $X$ and $Y$ are homeomorphic if there exists a bijection $h : X \rightarrow Y$ s. t. $h$ and $h^{-1}$ are continuous.

- $X, Y \subset \mathbb{R}^d$ are ambient isotopic if there exists a continuous map $F : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}^d$ s. t. $F(., 0) = Id_{\mathbb{R}^d}$, $F(X, 1) = Y$ and $\forall t \in [0, 1]$, $F(., t)$ is an homeomorphism of $\mathbb{R}^d$. 
Comparing topological spaces

Homotopy, homotopy type

- Two maps $f_0 : X \rightarrow Y$ and $f_1 : X \rightarrow Y$ are homotopic if there exists a continuous map $H : [0,1] \times X \rightarrow Y$ s. t. $\forall x \in X$, $H(0,x) = f_0(x)$ and $H(1,x) = f_1(x)$.

- $X$ and $Y$ have the same homotopy type (or are homotopy equivalent) if there exists continuous maps $f : X \rightarrow Y$ and $g : Y \rightarrow X$ s. t. $g \circ f$ is homotopic to $Id_X$ and $f \circ g$ is homotopic to $Id_Y$. 
Comparing topological spaces

Homotopy, homotopy type

If $X \subset Y$ and if there exists a continuous map $H : [0, 1] \times X \to X$ s.t.:

i) $\forall x \in X, \ H(0, x) = x$,

ii) $\forall x \in X, \ H(1, x) \in Y$

iii) $\forall y \in Y, \ \forall t \in [0, 1], \ H(t, y) \in Y$,

then $X$ and $Y$ are homotopy equivalent. If one replaces condition iii) by $\forall y \in Y, \ \forall t \in [0, 1], \ H(t, y) = y$ then $H$ is a deformation retract of $X$ onto $Y$. 

\[ f_0(x) = x \]

\[ f_t(x) = (1-t)x \]

\[ f_1(x) = 0 \]
Given a set $P = \{p_0, \ldots, p_k\} \subset \mathbb{R}^d$ of $k + 1$ affinely independent points, the $k$-dimensional simplex $\sigma$, or $k$-simplex for short, spanned by $P$ is the set of convex combinations

$$\sum_{i=0}^{k} \lambda_i p_i, \quad \text{with} \quad \sum_{i=0}^{k} \lambda_i = 1 \quad \text{and} \quad \lambda_i \geq 0.$$ 

The points $p_0, \ldots, p_k$ are called the vertices of $\sigma$. 

0-simplex: vertex  
1-simplex: edge  
2-simplex: triangle  
3-simplex: tetrahedron  

etc...
A (finite) simplicial complex $K$ in $\mathbb{R}^d$ is a (finite) collection of simplices such that:

1. any face of a simplex of $K$ is a simplex of $K$,
2. the intersection of any two simplices of $K$ is either empty or a common face of both.

The underlying space of $K$, denoted by $|K| \subset \mathbb{R}^d$ is the union of the simplices of $K$. 
Abstract simplicial complexes

Let \( P = \{p_1, \ldots, p_n\} \) be a (finite) set. An abstract simplicial complex \( K \) with vertex set \( P \) is a set of subsets of \( P \) satisfying the two conditions:

1. The elements of \( P \) belong to \( K \).
2. If \( \tau \in K \) and \( \sigma \subseteq \tau \), then \( \sigma \in K \).

The elements of \( K \) are the simplices.

Let \( \{e_1, \ldots, e_n\} \) a basis of \( \mathbb{R}^n \). “The” geometric realization of \( K \) is the (geometric) subcomplex \( |K| \) of the simplex spanned by \( e_1, \ldots, e_n \) such that:

\[
[e_{i_0} \ldots e_{i_k}] \in |K| \text{ iff } \{p_{i_0}, \ldots, p_{i_k}\} \in K
\]

\( |K| \) is a topological space (subspace of an Euclidean space)!
Abstract simplicial complexes

Let $P = \{p_1, \cdots, p_n\}$ be a (finite) set. An abstract simplicial complex $K$ with vertex set $P$ is a set of subsets of $P$ satisfying the two conditions:

1. The elements of $P$ belong to $K$.
2. If $\tau \in K$ and $\sigma \subseteq \tau$, then $\sigma \in K$.

The elements of $K$ are the simplices.

**IMPORTANT**

Simplicial complexes can be seen at the same time as geometric/topological spaces (good for top./geom. inference) and as combinatorial objects (abstract simplicial complexes, good for computations).
Covers and nerves

An open cover of a topological space $X$ is a collection $\mathcal{U} = (U_i)_{i \in I}$ of open subsets $U_i \subseteq X$, $i \in I$ where $I$ is a set, such that $X = \bigcup_{i \in I} U_i$.

Given a cover of a topological space $X$, $\mathcal{U} = (U_i)_{i \in I}$, its nerve is the abstract simplicial complex $C(\mathcal{U})$ whose vertex set is $\mathcal{U}$ and such that

$$\sigma = [U_{i_0}, U_{i_1}, \cdots, U_{i_k}] \in C(\mathcal{U}) \text{ if and only if } \bigcap_{j=0}^{k} U_{i_j} \neq \emptyset.$$
The Nerve Theorem:
Let $\mathcal{U} = (U_i)_{i \in I}$ be a finite open cover of a subset $X$ of $\mathbb{R}^d$ such that any intersection of the $U_i$'s is either empty or contractible. Then $X$ and $C(\mathcal{U})$ are homotopy equivalent.

For non-experts, you can replace:
- “contractible” by “convex”,
- “are homotopy equivalent” by ”have many topological invariants in common”.
Building interesting covers and nerves

Two directions:

1. Covering data by balls:
   → distance functions frameworks, persistence-based signatures,...
   → geometric inference, provide a framework to establish various theoretical results in TDA.

2. Using a function defined on the data:
   → the Mapper algorithm
   → exploratory data analysis and visualization
Covers and nerves for exploratory data analysis.
Let $f : X \to \mathbb{R}$ (or $\mathbb{R}^d$) a continuous function where $X$ is a topological space and let $\mathcal{U} = (U_i)_{i \in I}$ be a cover of $\mathbb{R}$ (or $\mathbb{R}^d$).

The collection of open sets $(f^{-1}(U_i))_{i \in I}$ is the pull back cover of $X$ induced by $(f, \mathcal{U})$. 
Take the connected components of the $f^{-1}(U_i), i \in I \to$ the refined pull back cover.

Take the nerve of the refined cover.

Warning: The nerve theorem does not apply in general!
The Mapper algorithm

**Input:**
- a data set $X$ with a metric or a dissimilarity measure,
- a function $f : X \to \mathbb{R}$ or $\mathbb{R}^d$,
- a cover $\mathcal{U}$ of $f(X)$.

1. for each $U \in \mathcal{U}$, decompose $f^{-1}(U)$ into clusters $C_{U,1}, \cdots, C_{U,k_U}$.
2. Compute the nerve of the cover of $X$ defined by the $C_{U,1}, \cdots, C_{U,k_U}, U \in \mathcal{U}$

**Output:** a simplicial complex, the nerve (often a graph for well-chosen covers → easy to visualize):
- a vertex $v_{U,i}$ for each cluster $C_{U,i}$,
- an edge between $v_{U,i}$ and $v_{U',j}$ iff $C_{U,i} \cap C_{U',j} \neq \emptyset$
The Mapper algorithm

Input:
- a data set $X$ with a metric or a dissimilarity measure,
- a function $f : X \rightarrow \mathbb{R}$ or $\mathbb{R}^d$, 
- a cover $\mathcal{U}$ of $f(X)$.

1. for each $U \in \mathcal{U}$, decompose $f^{-1}(U)$ into clusters $C_{U,1}, \cdots, C_{U,k_U}$.
2. Compute the nerve of the cover of $X$ defined by the $C_{U,1}, \cdots, C_{U,k_U}$, $U \in \mathcal{U}$

Output: a simplicial complex, the nerve (often a graph for well-chosen covers $\rightarrow$ easy to visualize):
- a vertex $v_{U,i}$ for each cluster $C_{U,i}$,
- an edge between $v_{U,i}$ and $v_{U',j}$ iff $C_{U,i} \cap C_{U',j} \neq \emptyset$

A very simple method but many choices to make!

Many (still open) theoretical questions!
Choice of lens/filter

\[ f : X \rightarrow \mathbb{R} \] is often called a lens or a filter.

Classical choices:

- density estimates
- centrality \( f(x) = \sum_{y \in X} d(x, y) \)
- excentricity \( f(x) = \max_{y \in X} d(x, y) \)
- PCA coordinates, NLDR coordinates,…
- Eigenfunctions of graph laplacians.
- Functions detecting anomalous behavior or outliers.
- Distance to a root point (filamentary structures reconstruction).
- Etc …
Choice of lens/filter

\( f : X \rightarrow \mathbb{R} \) is often called a lens or a filter.

Classical choices:

- density estimates
- centrality \( f(x) = \sum_{y \in X} d(x, y) \)
- excentricity \( f(x) = \max_{y \in X} d(x, y) \)
- PCA coordinates, NLDR coordinates, ...
- Eigenfunctions of graph laplacians.
- Functions detecting anomalous behavior or outliers.
- Distance to a root point (filamentary structures reconstruction).
- Etc ...
**Choice of lens/filter**

\( f : X \rightarrow \mathbb{R} \) is often called a **lens** or a **filter**.

May reveal some ambiguity in the use of non linear dimensionality reduction (NLDR) methods.

**Classical choices:**

- density estimates
- centrality \( f(x) = \sum_{y \in X} d(x, y) \)
- excentricity \( f(x) = \max_{y \in X} d(x, y) \)
- PCA coordinates, NLDR coordinates,…
- Eigenfunctions of graph laplacians.
- Functions detecting anomalous behavior or outliers.
- Distance to a root point (filamentary structures reconstruction).
- Etc …
The resolution $r$ is the maximum diameter of an interval in $\mathcal{U}$. The resolution may also be replaced by a number $N$ of intervals in the cover.

The gain $g$ is the percentage of overlap between intervals (when they overlap).

**Intuition:**
- small $r$ (large $N$) $\rightarrow$ finer resolution, more nodes.
- large $r$ (small $N$) $\rightarrow$ rougher resolution, less nodes.
- small $g$ $\rightarrow$ less connectivity.
- large $g$ $\rightarrow$ more connectivity (the dimensionality of the nerve increases).
Choice of covers (case of $\mathbb{R}$)

The **resolution** $r$ is the maximum diameter of an interval in $\mathcal{U}$. The resolution may also be replaced by a number $N$ of intervals in the cover. The **gain** $g$ is the percentage of overlap between intervals (when they overlap).

The resolution $r$ is the maximum diameter of an interval in $\mathcal{U}$. The resolution may also be replaced by a number $N$ of intervals in the cover. The gain $g$ is the percentage of overlap between intervals (when they overlap).

**Intuition:**
- small $r$ (large $N$) $\rightarrow$ finer resolution, more nodes.
- large $r$ (small $N$) $\rightarrow$ rougher resolution, less nodes.
- small $g$ $\rightarrow$ less connectivity.
- large $g$ $\rightarrow$ more connectivity (the dimensionality of the nerve increases).

**Major warning:** the output of Mapper is very sensitive to the choice of the parameters (see practical classes).

Not a well-understood phenomenon
Choice of clusters

2 strategies:
Choice of clusters

2 strategies:

1. Take the connected components of the subgraph spanned by the vertices in the bin $f^{-1}(U)$.

2. Build a neighboring graph (kNN, Rips,...). In general, need to select a global parameter, such as number of neighbors for kNN, radius for Rips, to build the graph: not adaptative.

Take the connected components of the subgraph spanned by the vertices in the bin $f^{-1}(U)$. 
Choice of clusters

2 strategies:

Clustering of each bin $f^{-1}(U)$ (using your favorite clustering algorithm)

More adaptive: the clustering parameters (or even the clustering algorithm) can be adapted to each bin.
Two “classical” applications of Mapper: clustering and feature selection

Clustering:

1. Build a Mapper graph/complex from the data,
2. Find interesting structures (loops, flares),
3. Use these structures to exhibit interesting clusters.
Two “classical” applications of Mapper: clustering and feature selection

**Clustering:**

1. Build a Mapper graph/complex from the data,
2. Find interesting structures (loops, flares),
3. Use these structures to exhibit interesting clusters.

**Some difficulties:**

**Choice of the parameters?**

**Done by hand...**

**Statistical relevance?**
Two “classical” applications of Mapper: clustering and feature selection

Clustering:

Example:
Data: conformations of molecules
Goal: detect different folding pathways

\( f : \) distance to folded/unfolded states
\( N = 8, \ g = 0.25 \)

Idea: 1 loop = 2 different pathways

Topological Methods for Exploring Low-density States in Biomolecular Folding Pathways, Yao et al., J. Chemical Physics, 2009
Two “classical” applications of Mapper: clustering and feature selection

Feature selection:

1. Build a Mapper graph/complex from the data,
2. Find interesting structures (loops, flares),
3. Select the features/variables that best discriminate the data in these structures.
Two “classical” applications of Mapper: clustering and feature selection

Feature selection:

Some difficulties:

Choice of the parameters?

1. Build a Mapper graph/complex from the data,
2. Find interesting structures (loops, flares),
3. Select the features/variables that best discriminate the data in these structures.

Done by hand...
Two “classical” applications of Mapper: clustering and feature selection

Feature selection:

Example:

Data: breast cancer patients that went through specific therapy.

Extracting insights from the shape of complex data using topology, Lum et al., Nature, 2013

\( f : \) eccentricity, \( N = 30, g = 0.33 \)

Goal: detect variables that influence survival after therapy in breast cancer patients
Reeb graph and Mapper

The output of the Mapper algorithm can be seen as a discretized version of the Reeb graph.

Equivalence relation:
\[ x \sim x' \text{ iff } x \text{ and } x' \text{ are in the same connected comp. of } f^{-1}(f(x)). \]

Reeb “graph”:
\[ G_f := X/ \sim \]

Warning:
- \( G_f \) is not always a graph (very specific conditions on \( X \) and \( f \)),
- No clear connection or convergence result relating the Mapper graph and the Reeb graph.
Reeb graph and Mapper

**Exercise:** What is the Mapper/Reeb graph of the height function on the trefoil knot?
Take-home messages

The Mapper algorithm:
1. local clustering guided by a function,
2. global connectivity relationships between clusters (covers and nerves).
→ other ways to combine local clustering, covers and nerves can be imagined!

The Mapper methods is an exploratory data analysis tool:
+ it has been shown to be very powerfull in various applications,
- but it usually does not come with theoretical guarantees.

Covers and nerves:
+ very interesting, simple and fruitfull ideas for topological data analysis,
+ many ideas and open questions to explore (in a statistical and data analysis perspective) from the theoretical point of view.
A few basic ideas about geometric inference: union of balls and distance functions
Union of balls and distance functions

Data set: a point cloud $P$ embedded in $\mathbb{R}^d$, sampled around a compact set $M$.

General idea:

1. Cover the data with union of balls of fixed radius centered on the data points.

2. Infer topological information about $M$ from (the nerve of) the union of balls centered on $P$. 

Union of balls and distance functions

Data set: a point cloud $P$ embedded in $\mathbb{R}^{d}$, sampled around a compact set $M$.

**General idea:**

1. Cover the data with union of balls of fixed radius centered on the data points.

2. Infer topological information about $M$ from (the nerve of) the union of balls centered on $P$.

Nerve theorem

Bridge the gap between continuous approximations of $K$ and combinatorial descriptions required by algorithms.
**Union of balls and distance functions**

Data set: a point cloud $P$ embedded in $\mathbb{R}^d$, sampled around a compact set $M$.

**General idea:**

1. Cover the data with union of balls of fixed radius centered on the data points.

2. Infer topological information about $M$ from (the nerve of) the union of balls centered on $P$.

Sublevel set of the **distance function** $d_P : \mathbb{R}^d \to \mathbb{R}_+$ is defined by

$$d_P(x) = \inf_{p \in P} \|x - p\|$$

→ Compare the topology/geometry of the offsets

$$M^r = d_M^{-1}([0, r]) \text{ and } P^r = d_P^{-1}([0, r])$$
Union of balls and distance functions

Data set: a point cloud $P$ embedded in $\mathbb{R}^d$, sampled around a compact set $M$.

General idea:

1. Cover the data with union of balls of fixed radius centered on the data points.

2. Infer topological information about $M$ from (the nerve of) the union of balls centered on $P$.

Sublevel set of the distance function $d_P : \mathbb{R}^d \to \mathbb{R}^+$ is defined by

$$d_P(x) = \inf_{p \in P} \|x - p\|$$

→ Compare the topology/geometry of the offsets

$$M^r = d_M^{-1}([0, r]) \text{ and } P^r = d_P^{-1}([0, r])$$

Regularity conditions? Sampling conditions?
The Hausdorff distance

The distance function to a compact $M \subset \mathbb{R}^d$, $d_M : \mathbb{R}^d \to \mathbb{R}_+$ is defined by

$$d_M(x) = \inf_{p \in M} \|x - p\|$$

The Hausdorff distance between two compact sets $M, M' \subset \mathbb{R}^d$:

$$d_H(M, M') = \sup_{x \in \mathbb{R}^d} |d_M(x) - d_{M'}(x)|$$
Medial axis and critical points

\[ \Gamma_M(x) = \{ y \in M : d_M(x) = \|x - y\| \} \]

The Medial axis of \( M \):

\[ \mathcal{M}(M) = \{ x \in \mathbb{R}^d : |\Gamma_M(x)| \geq 2 \} \]

\( x \in \mathbb{R}^d \) is a critical point of \( d_M \) iff \( x \) is contained in the convex hull of \( \Gamma_M(x) \).

**Theorem:** [Grove, Cheeger,...] Let \( M \subset \mathbb{R}^d \) be a compact set.

- if \( r \) is a regular value of \( d_M \), then \( d_M^{-1}(r) \) is a topological submanifold of \( \mathbb{R}^d \) of codim 1.
- Let \( 0 < r_1 < r_2 \) be such that \([r_1, r_2]\) does not contain any critical value of \( d_M \).
  Then all the level sets \( d_M^{-1}(r), r \in [r_1, r_2] \) are isotopic and
  \[ M^{r_2} \setminus M^{r_1} = \{ x \in \mathbb{R}^d : r_1 < d_M(x) \leq r_2 \} \]
  is homeomorphic to \( d_M^{-1}(r_1) \times (r_1, r_2] \).
The reach of $M$, $\tau(M)$, is the smallest distance from $M(M)$ to $M$:

$$\tau(M) = \inf_{y \in M(M)} d_M(y)$$

The weak feature size of $M$, $wfs(M)$, is the smallest distance from the set of critical points of $d_M$ to $M$:

$$wfs(M) = \inf\{d_M(y) : y \in \mathbb{R}^d \setminus M \text{ and } y \text{ crit. point of } d_M\}$$
Reach, $\mu$-reach and geometric inference
(Not developed in this course - just an example of result)

"Theorem:" Let $M \subset \mathbb{R}^d$ be such that $\tau = \tau(M) > 0$ and let $P \subset \mathbb{R}^d$ be such that $d_H(M, P) < c\tau$ for some (explicit) constant $c$. Then, for well-chosen (and explicit) $r$, $P^r$, and thus its nerve, is homotopy equivalent to $M$.

More generally, for compact sets with positive $\mu$-reach ($\text{wfs}(M) \leq r_\mu(M) \leq \tau(M)$):

**Topological/geometric properties of the offsets of $K$ are stable with respect to Hausdorff approximation:**

1. Topological stability of the offsets (CCSL'06, NSW'06).
2. Approximate normal cones (CCSL’08).
3. Boundary measures (CCSM’07), curvature measures (CCSLT’09), Voronoi covariance measures (GMO’09).
The probabilistic setting

Let \( M \subset \mathbb{R}^d \) be a \( k \)-dim compact submanifold with positive reach \( r_1(M) \geq \tau > 0 \).

Let \( \mu \) be a probability measure such that \( \text{Supp}(\mu) = M \) which is \((a,k)\)-standard: there exists \( r_0 \geq \tau/8 > 0 \) such that for any \( x \in M \), \( \mu(B(x,r)) \geq ar^k \).

Let \( X = \{x_1, \cdots, x_n\} \subset \mathbb{R}^d \) be \( n \) points i.i.d. sampled according to \( \mu \).

**Goal:** Upper bound \( P(X^r \not\sim M) \) where \( \sim \) denotes the homotopy equivalence.

Connection to support estimation problems: it is enough to bound \( P(d_H(X, M) > \varepsilon) \).
Minimax risk

Let $Q = Q(d, k, \tau, a)$ be the family of probability measures on $\mathbb{R}^d$ such that for any $\mu \in Q$:
- $\text{Supp}(\mu)$ is a compact $k$-dimensional manifold with positive reach larger than $\tau$;
- $\mu$ is $(a, k)$-standard.

Given $\mu \in Q$, $\text{Supp}(\mu) = M$, denote by $\hat{M}$ any homotopy type estimator of $M$ that takes as input $n$-uples of points from $M$ and outputs a set whose homotopy type “estimates” the homotopy type of $M$ (e.g. a union of balls).

$$R_n = \inf_{\hat{M}} \sup_{Q \in Q} Q^n(\hat{M} \neq M)$$

**Theorem:** There exist constants $C_a, C_a', C_a'' > 0$ such that

$$\frac{1}{8} \exp(-nC_a \tau^k) \leq R_n \leq C_a' \frac{1}{\tau^k} \exp(-nC_a'' \tau^k)$$
Minimax risk

Let $Q = Q(d, k, \tau, a)$ be the family of probability measures on $\mathbb{R}^d$ such that for any $\mu \in Q$:
- $\text{Supp}(\mu)$ is a compact $k$-dimensional manifold with positive reach larger than $\tau$;
- $\mu$ is $(a, k)$-standard.

Given $\mu \in Q$, $\text{Supp}(\mu) = M$, denote by $\hat{M}$ any homotopy type estimator of $M$ that takes as input $n$-uples of points from $M$ and outputs a set whose homotopy type “estimates” the homotopy type of $M$ (e.g. a union of balls).

\[
R_n = \inf_{\hat{M}} \sup_{Q \in \mathcal{Q}} Q^n(\hat{M} \not\sim M)
\]

**Theorem:** There exist constants $C_a, C'_a, C''_a > 0$ such that

\[
\frac{1}{8} \exp(-nC_a \tau^k) \leq R_n \leq C'_a \frac{1}{\tau^k} \exp(-nC''_a \tau^k)
\]
More details on geometric inference and minimax convergence rates
Reconstruction theorem: (Weak version)
Let $K \subset \mathbb{R}^d$ be a compact set s.t. $r_\mu = r_\mu(K) > 0$ for some $\mu > 0$. Let $K' \subset \mathbb{R}^d$ be such that $d_H(K, K') = \varepsilon < \kappa r_\mu(K)$ with $\kappa < \frac{\mu^2}{5\mu^2 + 12}$ Then for any $d, d'$ s.t.

$$0 < d < \text{wfs}(K) \quad \text{and} \quad \frac{4\varepsilon}{\mu^2} \leq d' < r_\mu - 3\varepsilon$$

the offsets $K'^d$ and $K^d$ are homotopy equivalent.
Reconstruction theorem: smooth case

Let $M \subset \mathbb{R}^d$ be a $k$-dimensional compact submanifold with positive reach $r_1(M) \geq \tau > 0$.

**Lemma:** for any $0 < r < \tau$, the offset $M^r$ deformation retracts on $M$. In particular, $M^r$ and $M$ are homotopy equivalent.

**Reconstruction theorem:**
Let $X \subset \mathbb{R}^d$ be a compact set such that $d_H(M, X) = \varepsilon < \frac{1}{17} \tau$ Then for any $r$ s.t. $4\varepsilon \leq r < \tau - 3\varepsilon$ the offset $X^r$ and $M$ are homotopy equivalent.

**Rem:** A more careful analysis leads to slightly better bounds [NSW 2008, CL 2008].
Let $M \subset \mathbb{R}^d$ be a $k$-dim compact submanifold with positive reach $r_1(M) \geq \tau > 0$.

Let $P$ be a probability measure such that $\text{Supp}(P) = M$ which is $(a, k)$-standard: there exists $r_0 \geq \tau/8 > 0$ such that for any $x \in M$, $P(B(x, r)) \geq ar^k$.

Let $X = \{x_1, \cdots, x_n\} \subset \mathbb{R}^d$ be $n$ points i.i.d. sampled according to $P$.

**Goal:** Upper bound $P(X^{r'} \not\cong M)$ where $\cong$ denotes the homotopy equivalence.
Lemma: Let \( \{A_i\}_{i=1}^{l} \) be a finite collection of measurable sets such that \( M \subset \bigcup_{i=1}^{l} A_i \) and \( P(A_i) > \alpha \) for some \( \alpha > 0 \).
Let \( X = \{x_1, \ldots, x_n\} \) a set of \( n \) i.i.d. points sampled according to \( P \).
Given \( \delta \in (0, 1) \), if \( n \geq \frac{1}{\alpha} \left( \log(l) + \log\left(\frac{1}{\delta}\right) \right) \) then with probability larger than \( 1 - \delta \), one has \( X \cap A_i \neq \emptyset \) for any \( i = 1, \ldots, l \).
In other words,
\[
P(X \cap A_i = \emptyset, \text{ for some } i = 1, \ldots, l) \leq 1 - e^{-n\alpha}.
\]

Proof:
- Let \( E_i \) be the event \( X \cap A_i = \emptyset \)
- \( P(E_i) = (1 - P(A_i))^n \leq (1 - \alpha)^n \)
- \( P(\bigcup E_i) \leq \sum P(E_i) \leq l(1 - \alpha)^n \)
- Use that \( (1 - \alpha) \leq e^{-\alpha} \)

Idea: Take \( A_i = B(x_i, r) \) and bound \( l \).
Covering and packing numbers

\[ C_M(r) = \text{minimum number of balls of radius } r \text{ needed to cover } M \]

\[ P_M(r) = \text{maximum number of balls of radius } r \text{ and center on } M \text{ that may be packed into } M \text{ without overlap in } M \]

\[ = \max \{ k : \exists y_1 \cdots y_k \in M \text{ s.t. } \forall i \neq j, B(y_i, r) \cap B(y_j, r) \cap M = \emptyset \}. \]

**Lemma:**

\[ P_M(2r) \leq C_M(2r) \leq P_M(r) \]

**Corollary:** for any \( r \leq 2r_0 \),

\[ C_M(r) \leq P_M\left(\frac{r}{2}\right) \leq \frac{2^d}{a} r^{-k} \]
An upper bound

1. As soon as \( d_H(X_n, M) < \frac{\tau}{8} \), one can recover the homotopy type of \( M \) from \( X^{r'}_n \) for well chosen \( r' \).

2. Let \( r = \frac{\tau}{8} \). Then
   \[
   C_M(r) \leq \frac{2^{4k}}{a} \left( \frac{1}{\tau} \right)^k
   \]

3. Let \( B_1, \cdots B_l \) a covering of \( M \) by balls of radius \( r \) with \( l \leq \frac{2^{4k}}{a} \left( \frac{1}{\tau} \right)^k \).
   For any \( i = 1, \cdots, l \), \( P(B_i) \geq \alpha = ar^k = \frac{a}{2^{3k}} \tau^k \).

4. Then
   \[
   P(d_H(X_n, M) > \frac{\tau}{8}) \leq le^{-n\alpha} \leq \frac{2^{4k}}{a} \frac{1}{\tau^k} e^{-n\frac{a}{2^{3k}} \tau^k}
   \]

Corollary: Let \( a' = \frac{a}{2^{3k}} \) and let \( r' \in \left( \frac{7\tau}{16}, \frac{5\tau}{8} \right) \). then
   \[
   P(X^{r'} \not\simeq M) \leq \frac{2^k}{a'} \frac{1}{\tau^k} e^{-na' \tau^k}
   \]

Rem: This bound only depends on \( a, k \) and \( \tau \).
Minimax risk

Let $Q = Q(d, k, \tau, a)$ be the family of probability measures on $\mathbb{R}^d$ such that for any $Q \in Q$:
- $\text{Supp}(Q)$ is a compact $k$-dimensional manifold with positive reach larger than $\tau$;
- $Q$ is $(a, k)$-standard.

Given $Q \in Q$, $\text{Supp}(Q) = M$, denote by $\hat{M}$ any homotopy type estimator of $M$ that takes as input $n$-uples of points from $M$ and outputs a set whose homotopy type “estimates” the homotopy type of $M$ (e.g. a union of balls).

$$R_n = \inf_{\hat{M}} \sup_{Q \in Q} Q^n(\hat{M} \not\sim M)$$

Proposition:

$$R_n \leq 2^k \frac{1}{a'} \frac{1}{\tau^k} \exp(-n a' \tau^k)$$
Minimax risk

Let $Q = Q(d, k, \tau, a)$ be the family of probability measures on $\mathbb{R}^d$ such that for any $Q \in Q$:
- $\text{Supp}(Q)$ is a compact $k$-dimensional manifold with positive reach larger than $\tau$;
- $Q$ is $(a, k)$-standard.

Given $Q \in Q$, $\text{Supp}(Q) = M$, denote by $\hat{M}$ any homotopy type estimator of $M$ that takes as input $n$-uples of points from $M$ and outputs a set whose homotopy type “estimates” the homotopy type of $M$ (e.g. a union of balls).

$$R_n = \inf_{\hat{M}} \sup_{Q \in Q} Q^n(\hat{M} \not\sim M)$$

Proposition:

$$R_n \leq \frac{2^k}{a'} \frac{1}{\tau^k} \exp(-na'\tau^k)$$

Lower bound for $R_n$?
Lemma: Let $Q$ be a set of probability distributions and let $\theta(Q)$ taking values in a metric space with metric $\rho$. Let $Q_1, Q_2 \in Q$ be any pair of distributions. Let $x_1 \cdots x_n$ be $n$ points i.i.d sampled from some $Q \in Q$. Then

$$\inf_{\hat{\theta}} \sup_{Q \in Q} \mathbb{E}_{Q^n}[\rho(\hat{\theta}(x_1, \cdots, x_n), \theta(Q))] \geq \frac{1}{4} \rho(\theta(Q_1), \theta(Q_2))(1 - TV(Q_1, Q_2))^{2n}$$

where $Q^n$ is the product measure and $TV(., .)$ is the total variation distance.
Lecam lemma

**Lemma:** Let $\mathcal{Q}$ be a set of probability distributions and let $\theta(\mathcal{Q})$ taking values in a metric space with metric $\rho$. Let $\mathcal{Q}_1, \mathcal{Q}_2 \in \mathcal{Q}$ be any pair of distributions. Let $x_1 \cdots x_n$ be $n$ points i.i.d sampled from some $\mathcal{Q} \in \mathcal{Q}$. Then

$$\inf_{\hat{\theta}} \sup_{\mathcal{Q} \in \mathcal{Q}} \mathbb{E}_{\mathcal{Q}^n} \left[ \rho(\hat{\theta}(x_1, \cdots, x_n), \theta(\mathcal{Q})) \right] \geq \frac{1}{4} \rho(\theta(\mathcal{Q}_1), \theta(\mathcal{Q}_2))(1 - TV(\mathcal{Q}_1, \mathcal{Q}_2))^{2n}$$

where $\mathcal{Q}^n$ is the product measure and $TV(., .)$ is the total variation distance.

In our case:
- $\mathcal{Q}$ as before
- metric space: the set of homotopy equivalent classes of compact subsets of $\mathbb{R}^d$ with $\rho(K, K') = 1$ if $K$ and $K'$ are not homotopy equivalent, and 0 otherwise.
Lower bound

\( M_1 = S^k(0, R) \) : a \( k \) dim. sphere of radius \( R > \tau \).

\( M_2 = M_1 \cup S^k(\tau) \) where \( S^k(\tau) \) is at distance at least \( 2\tau \) from \( M_1 \).

\( v_k = \text{vol}(S^k(0, 1)) \)

\( Q_1: \) unif density (w.r.t. \( k \)-vol) \( \rightarrow f_1 = \frac{1}{v_k R^k} \)

\( Q_2: \) unif density (w.r.t. \( k \)-vol) \( \rightarrow f_2 = \frac{1}{v_k (R^k + \tau^k)} \)

\( TV(Q_1, Q_2) = Q_2(M_2 \backslash M_1) - Q_1(M_2 \backslash M_1) \)

\[ = f_2 v_k \tau^k - 0 = \frac{\tau^d}{R^k + \tau^k} \leq C_R \tau^k \]

So, \( (1 - TV(Q_1, Q_2))^2n \geq (1 - C_a \tau^k)^2n \)

\[ R_n \geq \frac{1}{8} (1 - C_a \tau^k)^{2n} \geq \frac{1}{8} \exp(-4C_a n \tau^k) \]
Affinity, total variation and Hellinger distances

Let $P$ and $Q$ be two ($\sigma$-finite, Borel) probability measures with density $p$ and $q$ with respect to any third measure that dominates both $P$ and $Q$.

**Definition:** Let $p \wedge q(x) = \min(p(x), q(x))$. The **affinity** between $P$ and $Q$ is

$$
\|P \wedge Q\| = \int p \wedge q = 1 - \frac{1}{2} \int |p - q|
$$

**Definition:** The **total variation** distance between $P$ and $Q$ is defined as

$$
TV(P, Q) = \sup_{A \text{ borel set}} |P(A) - Q(A)|
= P(G) - Q(G) \text{ where } G = \{x : p(x) \geq q(x)\}
= 1 - \int p \wedge q = 1 - \|P \wedge Q\|
$$

**Definition:** The **Hellinger distance** between $P$ and $Q$ is defined by

$$
h^2(P, Q) = \int (\sqrt{p} - \sqrt{q})^2 = 2(1 - \int \sqrt{pq})
$$
Proof of Le Cam lemma

Let $\hat{\theta}$ and $n$ be fixed.

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_Q^n (\rho(\hat{\theta}, \theta(Q))) \geq \frac{1}{2} \left[ \mathbb{E}_{Q_1^n} (\rho(\hat{\theta}, \theta(Q_1))) + \mathbb{E}_{Q_2^n} (\rho(\hat{\theta}, \theta(Q_2))) \right]$$
Proof of Le Cam lemma

Let $\hat{\theta}$ and $n$ be fixed.

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_Q^n (\rho(\hat{\theta}, \theta(Q))) \geq \frac{1}{2} \left[ \mathbb{E}_{Q_1}^n (\rho(\hat{\theta}, \theta(Q_1))) + \mathbb{E}_{Q_2}^n (\rho(\hat{\theta}, \theta(Q_2))) \right]$$

Let $\mu$ be a measure dominating $Q_1$ and $Q_2$.

$$A = \int \rho(\hat{\theta}, \theta(Q_1))q_{1,n}d\mu^n + \int \rho(\hat{\theta}, \theta(Q_2))q_{2,n}d\mu^n$$
Proof of Le Cam lemma

Let $\hat{\theta}$ and $n$ be fixed.

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_Q^n (\rho(\hat{\theta}, \theta(Q))) \geq \frac{1}{2} \left[ \mathbb{E}_{Q_1}^n (\rho(\hat{\theta}, \theta(Q_1))) + \mathbb{E}_{Q_2}^n (\rho(\hat{\theta}, \theta(Q_2))) \right]$$

Let $\mu$ be a measure dominating $Q_1$ and $Q_2$.

$$A = \int \rho(\hat{\theta}, \theta(Q_1)) q_{1,n} d\mu^n + \int \rho(\hat{\theta}, \theta(Q_2)) q_{2,n} d\mu^n$$

But $\rho(\hat{\theta}, \theta(Q_1)) + \rho(\hat{\theta}, \theta(Q_2)) \geq \rho(\theta(Q_1), \theta(Q_2))$
Proof of Le Cam lemma

Let $\hat{\theta}$ and $n$ be fixed.

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_Q^n (\rho(\hat{\theta}, \theta(Q))) \geq \frac{1}{2} \left[ \mathbb{E}_{Q_1^n} (\rho(\hat{\theta}, \theta(Q_1))) + \mathbb{E}_{Q_2^n} (\rho(\hat{\theta}, \theta(Q_2))) \right]$$

Let $\mu$ be a measure dominating $Q_1$ and $Q_2$.

$$A = \int \rho(\hat{\theta}, \theta(Q_1)) q_{1,n} d\mu^n + \int \rho(\hat{\theta}, \theta(Q_2)) q_{2,n} d\mu^n$$

But $\rho(\hat{\theta}, \theta(Q_1)) + \rho(\hat{\theta}, \theta(Q_2)) \geq \rho(\theta(Q_1), \theta(Q_2))$

$$A \geq \rho(\theta(Q_1), \theta(Q_2)) \int q_{1,n} \wedge q_{2,n} d\mu^n = \rho(\theta(Q_1), \theta(Q_2)) \| Q_1^n \wedge Q_2^n \|$$
Proof of Le Cam lemma

Let \( \hat{\theta} \) and \( n \) be fixed.

\[
\sup_{Q \in \mathcal{Q}} \mathbb{E}_Q^n (\rho(\hat{\theta}, \theta(Q))) \geq \frac{1}{2} \left[ \mathbb{E}_{Q_1}^n (\rho(\hat{\theta}, \theta(Q_1))) + \mathbb{E}_{Q_2}^n (\rho(\hat{\theta}, \theta(Q_2))) \right]
\]

Let \( \mu \) be a measure dominating \( Q_1 \) and \( Q_2 \).

\[
A = \int \rho(\hat{\theta}, \theta(Q_1)) q_1, n d\mu^n + \int \rho(\hat{\theta}, \theta(Q_2)) q_2, n d\mu^n
\]

But \( \rho(\hat{\theta}, \theta(Q_1)) + \rho(\hat{\theta}, \theta(Q_2)) \geq \rho(\theta(Q_1), \theta(Q_2)) \)

\[
A \geq \rho(\theta(Q_1), \theta(Q_2)) \int q_1, n \wedge q_2, n d\mu^n = \rho(\theta(Q_1), \theta(Q_2)) \| Q_1^n \wedge Q_2^n \|
\]

Lemma:

\[
\| Q_1^n \wedge Q_2^n \| \geq \frac{1}{2} \left( 1 - \frac{1}{2} \int |q_1 - q_2| \right)^{2n} = \frac{1}{2} \| Q_1 \wedge Q_2 \|^{2n}
\]
Proof of Le Cam lemma

Lemma:
\[ \| Q_1^n \land Q_2^n \| \geq \frac{1}{2} \left( 1 - \frac{1}{2} \int |q_1 - q_2| \right)^{2n} = \frac{1}{2} \| Q_1 \land Q_2 \|^{2n} \]

Proof:

Claim A: \( h^2(P, Q) \leq \int |p - q| = l_1(P, Q) \)

Claim B: \( h^2(P^n, Q^n) = 2 \left( 1 - [1 - \frac{h^2(P, Q)}{2}]^n \right) \)

Claim C: \( \left( 1 - \frac{h^2(P, Q)}{2} \right)^2 \leq 2 \| P \land Q \| \)

\[
\| Q_1^n \land Q_2^n \| \geq \frac{1}{2} \left( 1 - \frac{h^2(Q_1^n, Q_2^n)}{2} \right)^2 \quad (C)
\]
\[
= \frac{1}{2} \left( 1 - \frac{h^2(Q_1, Q_2)}{2} \right)^{2n} \quad (B)
\]
\[
\geq \frac{1}{2} \left( 1 - \frac{l_1(Q_1, Q_2)}{2} \right)^{2n} = \frac{1}{2} \| Q_1 \land Q_2 \|^{2n} \quad (A)
\]
References


