The Topology of Random Spaces

Elizabeth Meckes

Case Western Reserve University

LDHD Summer School
SAMSI
August, 2013

Statistical Topology

> I predict a new subject of statistical topology. Rather than count the number of holes, Betti numbers, etc., one will be more interested in the distribution of such objects on noncompact manifolds as one goes out to infinity.
>
> – Isadore Singer, 2004

Why study the topology of random spaces?

Why study the topology of random spaces?

- As a null hypothesis for statistical topology/to get a sense of what happens “generically”.

Why study the topology of random spaces?

- As a null hypothesis for statistical topology/to get a sense of what happens “generically”.
- For use in topological data analysis; in particular, to understand high-dimensional point cloud data via topological features.

Why study the topology of random spaces?

- As a null hypothesis for statistical topology/to get a sense of what happens “generically”.
- For use in topological data analysis; in particular, to understand high-dimensional point cloud data via topological features.
- Manifold learning.

Why study the topology of random spaces?

- As a null hypothesis for statistical topology/to get a sense of what happens "generically".
- For use in topological data analysis; in particular, to understand high-dimensional point cloud data via topological features.
- Manifold learning.
- For existence proofs via the probabilistic method.

The topology of data? Huh?

The nerve lemma and the Čech complex

Consider a collection of points $\mathcal{P}$. A natural way to make sense of “the topology” of $\mathcal{P}$ is to consider

$$\mathcal{U}_r(\mathcal{P}) := \cup_{p \in \mathcal{P}} B_r(p),$$

as a parametrized family of spaces for $r \in (0, \infty)$.

Consider a collection of points $\mathcal{P}$. A natural way to make sense of "the topology" of $\mathcal{P}$ is to consider

$$\mathcal{U}_r(\mathcal{P}) := \cup_{p \in \mathcal{P}} B_r(p),$$

as a parametrized family of spaces for $r \in (0, \infty)$.

The Čech complex $\mathcal{C}_r(\mathcal{P})$ on $\mathcal{P}$ is a simplicial complex with 0-skeleton $\mathcal{P}$ and faces included or not depending on the distances between the points of $\mathcal{P}$.

The topology of data? Huh?

The nerve lemma and the Čech complex

Consider a collection of points $\mathcal{P}$. A natural way to make sense of "the topology" of $\mathcal{P}$ is to consider

$$\mathcal{U}_r(\mathcal{P}) := \cup_{p \in \mathcal{P}} B_r(p),$$

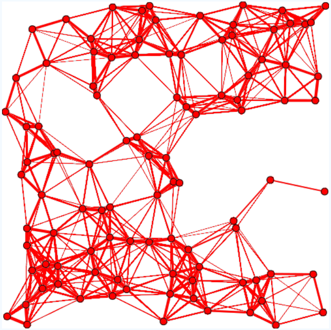
as a parametrized family of spaces for $r \in (0, \infty)$.

The Čech complex $\mathcal{C}_r(\mathcal{P})$ on $\mathcal{P}$ is a simplicial complex with 0-skeleton $\mathcal{P}$ and faces included or not depending on the distances between the points of $\mathcal{P}$.

The Nerve Lemma: The homology of $\mathcal{U}_r(\mathcal{P})$ and $\mathcal{C}_r(\mathcal{P})$ are the same.

The random Čech complex

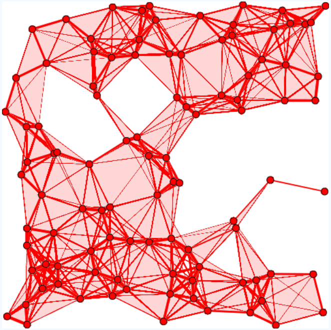
The random Čech complex



A geometric random graph

- First construct a random geometric graph: start with a random point process, and add an edge between v and w if $B_r(v) \cap B_r(w) \neq \emptyset$.

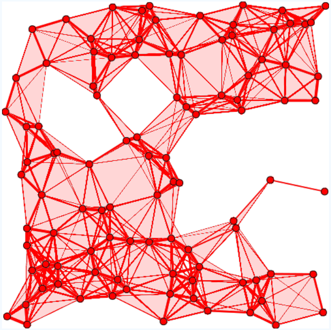
The random Čech complex



A random Čech complex

- First construct a random geometric graph: start with a random point process, and add an edge between v and w if $B_r(v) \cap B_r(w) \neq \emptyset$.
- Continue in higher dimensions: if $B_r(v) \cap B_r(w) \cap B_r(z) \neq \emptyset$, fill in the triangle with vertices v, w, z.

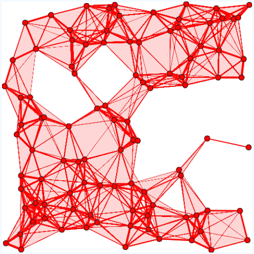
The random Čech complex



A random Čech complex

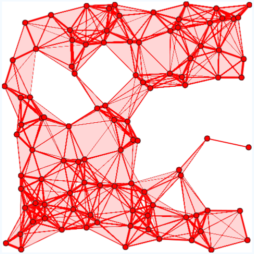
- First construct a random geometric graph: start with a random point process, and add an edge between v and w if $B_r(v) \cap B_r(w) \neq \emptyset$.
- Continue in higher dimensions: if $B_r(v) \cap B_r(w) \cap B_r(z) \neq \emptyset$, fill in the triangle with vertices v, w, z.
- And so on...

More specifics about the construction



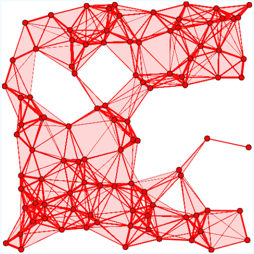
More specifics about the construction

- Let f be a bounded density on $\mathbb{R}^d$. Choose n points $\{X_1, \ldots, X_n\}$ independently according to f to be the vertices of the complex.



More specifics about the construction

- Let f be a bounded density on $\mathbb{R}^d$. Choose n points $\{X_1, \ldots, X_n\}$ independently according to f to be the vertices of the complex.
- Let r_n be the connectivity threshhold as described on the last slide.

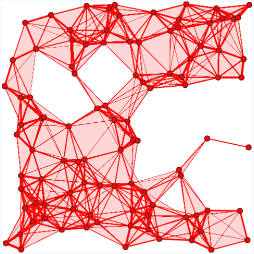


More specifics about the construction

- Let f be a bounded density on $\mathbb{R}^d$. Choose n points $\{X_1, \ldots, X_n\}$ independently according to f to be the vertices of the complex.

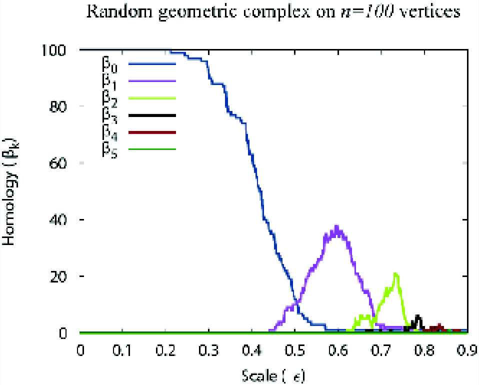
- Let r_n be the connectivity threshhold as described on the last slide.

- We study the random Čech complex $\mathcal{C} = \mathcal{C}(X_1, \ldots, X_n)$.



Realizations of Betti numbers

Realizations of Betti numbers



The Betti numbers of the random Vietoris–Rips complex vs. r_n, with $n = 100$. *Figure courtesy of Afra Zomorodian*.

The three regimes

The behavior of $\mathcal{C} = \mathcal{C}(X_1, \ldots, X_n)$ is qualitatively different depending on the asymptotics of the quantity nr_n^d, as follows:

The three regimes

The behavior of $\mathcal{C} = \mathcal{C}(X_1, \ldots, X_n)$ is qualitatively different depending on the asymptotics of the quantity nr_n^d, as follows:

- $nr_n^d \xrightarrow{n\to\infty} 0$: The sparse or sub-critical regime

The three regimes

The behavior of $\mathcal{C} = \mathcal{C}(X_1, \ldots, X_n)$ is qualitatively different depending on the asymptotics of the quantity nr_n^d, as follows:

- $nr_n^d \xrightarrow{n\to\infty} 0$: The sparse or sub-critical regime
- $nr_n^d \xrightarrow{n\to\infty} \alpha \in (0,\infty)$: The critical regime

The three regimes

The behavior of $\mathcal{C} = \mathcal{C}(X_1, \ldots, X_n)$ is qualitatively different depending on the asymptotics of the quantity nr_n^d, as follows:

- $nr_n^d \xrightarrow{n\to\infty} 0$: The sparse or sub-critical regime
- $nr_n^d \xrightarrow{n\to\infty} \alpha \in (0, \infty)$: The critical regime
- $nr_n^d \xrightarrow{n\to\infty} \infty$: The super-critical regime

Expected Betti numbers in the sparse regime

Expected Betti numbers in the sparse regime

K-M showed:

For $0 \leq k \leq d-1$, there is a constant μ depending only on f and k such that if $nr_n^d \xrightarrow{n\to\infty} 0$,

$$\frac{\mathbb{E}\left[\beta_k(\mathcal{C})\right]}{n^{k+2} r_n^{d(k+1)}} \longrightarrow \frac{\mu}{(k+1)!} \qquad \text{as} \qquad n \to \infty.$$

Expected Betti numbers in the sparse regime

K-M showed:

For $0 \leq k \leq d-1$, there is a constant μ depending only on f and k such that if $nr_n^d \xrightarrow{n \to \infty} 0$,

$$\frac{\mathbb{E}\left[\beta_k(\mathcal{C})\right]}{n^{k+2} r_n^{d(k+1)}} \longrightarrow \frac{\mu}{(k+1)!} \qquad \text{as} \qquad n \to \infty.$$

Expected Betti numbers in the sparse regime

K-M showed:

For $0 \leq k \leq d-1$, there is a constant μ depending only on f and k such that if $nr_n^d \xrightarrow{n\to\infty} 0$,

$$\frac{\mathbb{E}\left[\beta_k(\mathcal{C})\right]}{n^{k+2}r_n^{d(k+1)}} \longrightarrow \frac{\mu}{(k+1)!} \qquad \text{as} \qquad n \to \infty.$$

In particular, for k fixed there are three subregimes in the sparse regime:

- $n^{k+2}r_n^{d(k+1)} \to 0$;
- $n^{k+2}r_n^{d(k+1)} \to \beta \in (0,\infty)$;
- $n^{k+2}r_n^{d(k+1)} \to \infty$.

Limiting distributions of Betti numbers

Theorem (Kahle/M.)

1. *If* $n^{k+2} r_n^{d(k+1)} \to 0$ *as* $n \to \infty$, *then*

$$\beta_k(\mathcal{C}(X_1, \ldots, X_n)) \to 0 \quad \textit{a.a.s. as } n \to \infty.$$

Limiting distributions of Betti numbers

Theorem (Kahle/M.)

1. *If* $n^{k+2} r_n^{d(k+1)} \to 0$ *as* $n \to \infty$*, then*

$$\beta_k(\mathcal{C}(X_1, \ldots, X_n)) \to 0 \quad \textit{a.a.s. as } n \to \infty.$$

2. *If* $n^{k+2} r_n^{d(k+1)} \to \alpha \in (0, \infty)$ *as* $n \to \infty$*, then*

$$d_{TV}(\beta_k(\mathcal{C}(X_1, \ldots, X_n)), Y) \leq c n r_n^d,$$

where Y *is a Poisson random variable with* $\mathbb{E}[Y] = \mathbb{E}[\beta_k]$ *and* c *is a constant depending only on* α*,* k *and* f*.*

Limiting distributions of Betti numbers

Theorem (Kahle/M.)

1. *If* $n^{k+2} r_n^{d(k+1)} \to 0$ *as* $n \to \infty$, *then*

$$\beta_k(\mathcal{C}(X_1, \ldots, X_n)) \to 0 \quad \textit{a.a.s. as } n \to \infty.$$

2. *If* $n^{k+2} r_n^{d(k+1)} \to \alpha \in (0, \infty)$ *as* $n \to \infty$, *then*

$$d_{TV}(\beta_k(\mathcal{C}(X_1, \ldots, X_n)), Y) \leq c n r_n^d,$$

   where Y *is a Poisson random variable with* $\mathbb{E}[Y] = \mathbb{E}[\beta_k]$ *and* c *is a constant depending only on* α, k *and* f.

3. *If* $n^{k+2} r_n^{d(k+1)} \to \infty$ *and* $n r_n^d \to 0$ *as* $n \to \infty$, *then*

$$\frac{\beta(\mathcal{C}(X_1, \ldots, X_n)) - \mathbb{E}[\beta(\mathcal{C}(X_1, \ldots, X_n))]}{\sqrt{\mathrm{Var}(\beta(\mathcal{C}(X_1, \ldots, X_n)))}} \Rightarrow \mathcal{N}(0, 1).$$

A different tack: the distance function

A different tack: the distance function

Let $\mathcal{P}$ be a set of points in $\mathbb{R}^d$, and consider the distance function $d_{\mathcal{P}} : \mathbb{R}^d \to [0, \infty)$ defined by

$$d_{\mathcal{P}}(x) := \min_{p \in \mathcal{P}} \|x - p\|.$$

A different tack: the distance function

Let $\mathcal{P}$ be a set of points in $\mathbb{R}^d$, and consider the distance function $d_{\mathcal{P}} : \mathbb{R}^d \to [0, \infty)$ defined by

$$d_{\mathcal{P}}(x) := \min_{p \in \mathcal{P}} \|x - p\|.$$

Note:

$$\left\{x \in \mathbb{R}^d \middle| d_{\mathcal{P}}(x) \leq r\right\} = \mathcal{U}_r(\mathcal{P});$$

that is, the topology we're interested in is contained in the sublevel sets of the distance function.

Morse Theory

Morse Theory

Let $f : \mathcal{M} \to \mathbb{R}$ be a *Morse function*; i.e., a smooth function on a closed manifold $\mathcal{M}$ such that all critical points of f are nondegenerate and all critical values of f are distinct.

Morse Theory

Let $f : \mathcal{M} \to \mathbb{R}$ be a *Morse function*; i.e., a smooth function on a closed manifold $\mathcal{M}$ such that all critical points of f are nondegenerate and all critical values of f are distinct.

Consider the sublevel sets

$$\mathcal{M}_r := \left\{ x \in \mathcal{M} \middle| f(x) \leq r \right\}.$$

Morse Theory

Let $f : \mathcal{M} \to \mathbb{R}$ be a *Morse function*; i.e., a smooth function on a closed manifold $\mathcal{M}$ such that all critical points of f are nondegenerate and all critical values of f are distinct.

Consider the sublevel sets

$$\mathcal{M}_r := \left\{ x \in \mathcal{M} \middle| f(x) \leq r \right\} .$$

If there are no critical levels between a and b, $\mathcal{M}_a$ and $\mathcal{M}_b$ are homotopy equivalent.

As you pass through each critical value, one of the Betti numbers changes by one – a hole is created or filled in.

Morse theory has been extended to apply to min-type functions, e.g., the distance function, so that the homology of $\mathcal{U}_r(\mathcal{P})$ for stochastic point process $\mathcal{P}$ can be studied through the random function $d_{\mathcal{P}}$ and its critical points.

Morse theory has been extended to apply to min-type functions, e.g., the distance function, so that the homology of $\mathcal{U}_r(\mathcal{P})$ for stochastic point process $\mathcal{P}$ can be studied through the random function $d_{\mathcal{P}}$ and its critical points.

The challenge is that Morse theory says that at a critical value of "Morse index k", β_k goes up by 1 or else β_{k-1} goes down by 1, but the definition of Morse index assumes smoothness that $d_{\mathcal{P}}$ lacks.

A new definition of criticality

Gershkovich and Rubinstein (via Bobrowski and Adler) give the following definition of a critical point of index k of $d_{\mathcal{P}}$:

A new definition of criticality

Gershkovich and Rubinstein (via Bobrowski and Adler) give the following definition of a critical point of index k of $d_{\mathcal{P}}$:

Definition

The critical points of index 0 of $d_{\mathcal{P}}$ are the global minima of $d_{\mathcal{P}}$; i.e., the points of $\mathcal{P}$ itself.

A new definition of criticality

Gershkovich and Rubinstein (via Bobrowski and Adler) give the following definition of a critical point of index k of $d_{\mathcal{P}}$:

Definition

The critical points of index 0 *of* $d_{\mathcal{P}}$ *are the global minima of* $d_{\mathcal{P}}$*; i.e., the points of* $\mathcal{P}$ *itself.*

A point $x \in \mathbb{R}^d$ *is a critical point of index* $k \in \{1, \ldots, d\}$ *of* $d_{\mathcal{P}}$ *if there is a subset* $\mathcal{Y} \subseteq \mathcal{P}$ *with* $|\mathcal{Y}| = k + 1$ *such that*

A new definition of criticality

Gershkovich and Rubinstein (via Bobrowski and Adler) give the following definition of a critical point of index k of $d_{\mathcal{P}}$:

Definition

The critical points of index 0 *of* $d_{\mathcal{P}}$ *are the global minima of* $d_{\mathcal{P}}$*; i.e., the points of* $\mathcal{P}$ *itself.*

A point $x \in \mathbb{R}^d$ *is a critical point of index* $k \in \{1, \ldots, d\}$ *of* $d_{\mathcal{P}}$ *if there is a subset* $\mathcal{Y} \subseteq \mathcal{P}$ *with* $|\mathcal{Y}| = k + 1$ *such that*

- $d_{\mathcal{P}}(x) = \|x - y\|$ *for all* $y \in \mathcal{Y}$ *and* $\|x - p\| > d_{\mathcal{P}}$ *for all* $p \in \mathcal{P} \setminus \mathcal{Y}$.

A new definition of criticality

Gershkovich and Rubinstein (via Bobrowski and Adler) give the following definition of a critical point of index k of $d_{\mathcal{P}}$:

Definition

The critical points of index 0 *of* $d_{\mathcal{P}}$ *are the global minima of* $d_{\mathcal{P}}$*; i.e., the points of* $\mathcal{P}$ *itself.*

A point $x \in \mathbb{R}^d$ *is a critical point of index* $k \in \{1, \ldots, d\}$ *of* $d_{\mathcal{P}}$ *if there is a subset* $\mathcal{Y} \subseteq \mathcal{P}$ *with* $|\mathcal{Y}| = k + 1$ *such that*

- $d_{\mathcal{P}}(x) = \|x - y\|$ *for all* $y \in \mathcal{Y}$ *and* $\|x - p\| > d_{\mathcal{P}}$ *for all* $p \in \mathcal{P} \setminus \mathcal{Y}$.
- *The points of* $\mathcal{Y}$ *are in general position.*

A new definition of criticality

Gershkovich and Rubinstein (via Bobrowski and Adler) give the following definition of a critical point of index k of $d_{\mathcal{P}}$:

Definition

The critical points of index 0 *of* $d_{\mathcal{P}}$ *are the global minima of* $d_{\mathcal{P}}$*; i.e., the points of* $\mathcal{P}$ *itself.*

A point $x \in \mathbb{R}^d$ *is a critical point of index* $k \in \{1, \ldots, d\}$ *of* $d_{\mathcal{P}}$ *if there is a subset* $\mathcal{Y} \subseteq \mathcal{P}$ *with* $|\mathcal{Y}| = k + 1$ *such that*

- $d_{\mathcal{P}}(x) = \|x - y\|$ *for all* $y \in \mathcal{Y}$ *and* $\|x - p\| > d_{\mathcal{P}}$ *for all* $p \in \mathcal{P} \setminus \mathcal{Y}$.
- *The points of* $\mathcal{Y}$ *are in general position.*
- $x \in \mathrm{conv}^{\circ}(\mathcal{Y})$.

Bobrowski–Adler

Limit Theory for the number of critical points

Bobrowski–Adler

Limit Theory for the number of critical points

Given n i.i.d. points $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, each distributed according to the density f on $\mathbb{R}^d$, and given a threshhold radius r_n, let $N_{k,n}$ be the number of critical points of index k of $\mathcal{X}_n$.

Bobrowski–Adler

Limit Theory for the number of critical points

Given n i.i.d. points $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, each distributed according to the density f on $\mathbb{R}^d$, and given a threshhold radius r_n, let $N_{k,n}$ be the number of critical points of index k of $\mathcal{X}_n$.

B–A prove limit theorems for $N_{k,n}$ in all three regimes (with an additional condition in the supercritical regime).

Bobrowski–Adler

Limit Theory for the number of critical points

Given n i.i.d. points $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, each distributed according to the density f on $\mathbb{R}^d$, and given a threshhold radius r_n, let $N_{k,n}$ be the number of critical points of index k of $\mathcal{X}_n$.

B–A prove limit theorems for $N_{k,n}$ in all three regimes (with an additional condition in the supercritical regime).

The theorems in the sparse regime are exactly analogous to the limit theorems for β_k in the sparse regime, but the fact that they get theorems in all regimes points to how much more powerful the Morse theoretic approach can be over the direct topological approach of K–M.

The Expected Euler Characteristic

The Expected Euler Characteristic

Theorem (Bobrowski–Adler)

Let χ_n be the Euler characteristic of $\mathcal{C}_{r_n}(\mathcal{X}_n)$. Then

$$\lim_{n\to\infty} \frac{\mathbb{E}[\chi_n]}{n} = \begin{cases} 1 & nr_n^d \to 0; \\ 1 + \sum_{k=1}^{d} (-1)^k \gamma_k(\lambda) & nr_n^d \to \lambda \in (0,\infty); \\ 0 & nr_n^d \to \infty; \end{cases}$$

here, $\gamma_k(\lambda)$ are (sort of) explicit constants depending only on f, k and λ.

The Expected Euler Characteristic

Theorem (Bobrowski–Adler)

Let χ_n *be the Euler characteristic of* $\mathcal{C}_{r_n}(\mathcal{X}_n)$*. Then*

$$\lim_{n\to\infty} \frac{\mathbb{E}[\chi_n]}{n} = \begin{cases} 1 & nr_n^d \to 0; \\ 1 + \sum_{k=1}^{d} (-1)^k \gamma_k(\lambda) & nr_n^d \to \lambda \in (0,\infty); \\ 0 & nr_n^d \to \infty; \end{cases}$$

here, $\gamma_k(\lambda)$ *are (sort of) explicit constants depending only on* f*,* k *and* λ*.*

Moreover, when $nr_n^d \geq C_f \log(n)$*, then* $\mathbb{E}[\chi_n] \to 1$*.*

Sampling from a density on a manifold

Bobrowski–Mukherjee consider the problem of n i.i.d. points $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, each distributed according to a density f on an m-dimensional manifold (embedded in some Euclidean space).

Sampling from a density on a manifold

Bobrowski–Mukherjee consider the problem of n i.i.d. points $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, each distributed according to a density f on an m-dimensional manifold (embedded in some Euclidean space).

Again, the behavior splits into three regimes:

- $nr_n^m \xrightarrow{n\to\infty} 0$: The sparse or sub-critical regime
- $nr_n^m \xrightarrow{n\to\infty} \lambda \in (0, \infty)$: The critical regime
- $nr_n^m \xrightarrow{n\to\infty} \infty$: The super-critical regime

The sub-critical and critical regimes

The sub-critical and critical regimes

Bobrowski–Mukherjee prove detailed limit theorems for the Betti numbers and the number of critical points of a given index in the sub-critical regime, analogous to those in the Euclidean case.

The sub-critical and critical regimes

Bobrowski–Mukherjee prove detailed limit theorems for the Betti numbers and the number of critical points of a given index in the sub-critical regime, analogous to those in the Euclidean case.

Also as in the Euclidean case, the critical regime is too highly connected to get easily at the Betti numbers. In particular, B–M prove that if $nr_n^m \to \lambda \in (0, \infty)$, then

$$0 < \liminf_{n\to\infty} \frac{\mathbb{E}\left[\beta_{k,n}\right]}{n} \leq \limsup_{n\to\infty} \frac{\mathbb{E}\left[\beta_{k,n}\right]}{n} < \infty,$$

for each $k \in \{1, \ldots, m-1\}$.

Manifold learning

Can we recover the topology of an unknown manifold $\mathcal{M}$ by studying the topology of $\mathcal{U}_r(n)$ for n and r chosen suitably?

Manifold learning

Can we recover the topology of an unknown manifold $\mathcal{M}$ by studying the topology of $\mathcal{U}_r(n)$ for n and r chosen suitably?

Niyogi–Smale–Weinberger: Probably.

Manifold learning

Can we recover the topology of an unknown manifold $\mathcal{M}$ by studying the topology of $\mathcal{U}_r(n)$ for n and r chosen suitably?

Niyogi–Smale–Weinberger: Probably.

Theorem (N–S–W)

Given a compact Riemannian submanifold $\mathcal{M}$ of $\mathbb{R}^N$ with condition number $1/\tau$ and $\epsilon \in (0, \tau/2)$, there are explicit constants

$$N = N(\tau, \epsilon, \mathrm{vol}(\mathcal{M})) \qquad \delta = \delta(\tau, \epsilon, \mathrm{vol}(\mathcal{M}))$$

such that if $\mathcal{X} = \{X_1, \ldots, X_n\}$ is a sample of n i.i.d. points chosen uniformly from $\mathcal{M}$ and $n \geq N$, then with probability at least $1 - \delta$,

$$\mathcal{U}_\epsilon(\mathcal{X}) \textit{ has the same homology as } \mathcal{M}.$$

Manifold learning

Theorem (Bobrowski–Mukherjee)

Consider i.i.d. points $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ *distributed according to a density* f *on a smooth closed manifold* $\mathcal{M}$ *with bounded curvature. Suppose further that*

$$\min_{x \in \mathcal{M}} f(x) > 0.$$

There is a constant C *depending only on* f *such that if* $nr_n^m \geq C \log(n)$, *then with probability one,*

$$\beta_{k,n} = \beta_k(\mathcal{M})$$

for each $k \in \{0, \ldots, m\}$ *eventually.*

Striking out in a different direction: Negatively associated stationary point processes

Striking out in a different direction: Negatively associated stationary point processes

The previous results are about complexes built over i.i.d. point processes; essentially the same reslts hold in Euclidean space for complexes built over (nonhomogenous) Poisson point processes.

Striking out in a different direction: Negatively associated stationary point processes

The previous results are about complexes built over i.i.d. point processes; essentially the same reslts hold in Euclidean space for complexes built over (nonhomogenous) Poisson point processes.

In recent work, Yogeshwaran–Adler consider complexes built over a more general class of stationary point processes $\mathcal{P}$ in $\mathbb{R}^d$; i.e., given a Borel subset $B \subseteq \mathbb{R}^d$,

$$\mathbb{E}\big[\#\{p \in \mathcal{P} \cap B\}\big] = \mathrm{vol}(B).$$

Negatively associated stationary point processes

Specifically, Y–A found asymptotic expectations of Betti numbers for negatively associated stationary point processes, and observed that the asymptotic orders are different than in the i.i.d. case.

Negatively associated stationary point processes

Specifically, Y–A found asymptotic expectations of Betti numbers for negatively associated stationary point processes, and observed that the asymptotic orders are different than in the i.i.d. case.

For example, for the i.i.d Euclidean case, when $nr_n^d \geq C \log(n)$, the Čech complex becomes trivial with high probability; on a manifold, the topology coincides with that of the manifold.

If a Čech complex is constructed over the Ginibre process in $\mathbb{R}^d$, the corresponding cut-off happens at $\log(n)^{\frac{d}{4}}$.

Recall: Betti numbers in the sparse regime in $\mathbb{R}^d$

Theorem (Kahle/M.)

1. *If* $n^{k+2} r_n^{d(k+1)} \to 0$ *as* $n \to \infty$*, then*

$$\beta_k(\mathcal{C}(X_1, \ldots, X_n)) \to 0 \quad \textit{a.a.s. as } n \to \infty.$$

2. *If* $n^{k+2} r_n^{d(k+1)} \to \alpha \in (0, \infty)$ *as* $n \to \infty$*, then*

$$d_{TV}(\beta_k(\mathcal{C}(X_1, \ldots, X_n)), Y) \le c n r_n^d,$$

where Y *is a Poisson random variable with* $\mathbb{E}[Y] = \mathbb{E}[\beta_k]$ *and* c *is a constant depending only on* α*,* k *and* f*.*

3. *If* $n^{k+2} r_n^{d(k+1)} \to \infty$ *and* $n r_n^d \to 0$ *as* $n \to \infty$*, then*

$$\frac{\beta(\mathcal{C}(X_1, \ldots, X_n)) - \mathbb{E}[\beta(\mathcal{C}(X_1, \ldots, X_n))]}{\sqrt{\mathrm{Var}(\beta(\mathcal{C}(X_1, \ldots, X_n)))}} \Rightarrow \mathcal{N}(0, 1).$$

Some preliminaries to the proof

The first idea is to bound β_k between two combinatorial random variables counting potential contributions to homology, and prove the same limit theorems for both.

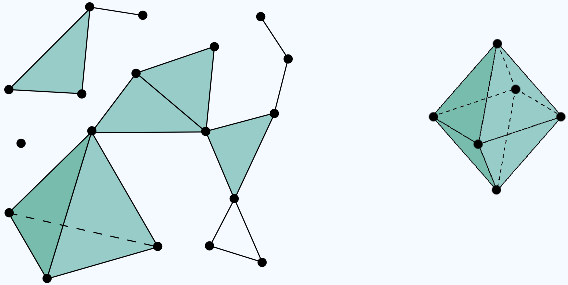
Some preliminaries to the proof

The first idea is to bound β_k between two combinatorial random variables counting potential contributions to homology, and prove the same limit theorems for both.

- Let $S_{n,k}$ be the number of empty k-simplices.
- Let $\widetilde{S}_{n,k}$, the number of *isolated* empty k-simplices.
- Let $Y_{n,k+1}$ denote the number of pairs $\{\sigma, \{u, v\}\}$, where $\sigma \subseteq \mathcal{C}(X_1, \ldots, X_n)$ is a k-simplex and u and v are distinct vertices of σ, each of which is connected to a different vertex outside σ.
- Let $Z_{n,k+1}$ denote the number of $\{\sigma, u\}$ in $\mathcal{C}(X_1, \ldots, X_n)$, where σ is a k-simplex and u is a vertex in σ with a path of length 2 outside σ attached.

It's not too hard to see that

$$\widetilde{S}_{n,k+1} \le \beta_k(\mathcal{C}(X_1, \ldots, X_n)) \le S_{n,k+1} + Y_{n,k+1} + Z_{n,k+1}.$$



$$\widetilde{S}_{n,k+1} \leq \beta_k(\mathcal{C}(X_1, \ldots, X_n)) \leq S_{n,k+1} + Y_{n,k+1} + Z_{n,k+1}$$

The idea is now to prove the same limit theorems for the upper and lower bounds using combinatorial techniques.

$$\widetilde{S}_{n,k+1} \leq \beta_k(\mathcal{C}(X_1, \ldots, X_n)) \leq S_{n,k+1} + Y_{n,k+1} + Z_{n,k+1}$$

The idea is now to prove the same limit theorems for the upper and lower bounds using combinatorial techniques.

What amounts to a rescaling argument with a little calculus shows that for a (not very explicit) constant μ,

$$\lim_{n\to\infty} \frac{\mathbb{E}[\widetilde{S}_{n,k+1}]}{n^{k+2} r_n^{d(k+1)}} = \lim_{n\to\infty} \frac{\mathbb{E}[S_{n,k+1} + Y_{n,k+1} + Z_{n,k+1}]}{n^{k+2} r_n^{d(k+1)}} = \frac{\mu}{(k+1)!}.$$

$$\widetilde{S}_{n,k+1} \leq \beta_k(\mathcal{C}(X_1, \ldots, X_n)) \leq S_{n,k+1} + Y_{n,k+1} + Z_{n,k+1}$$

The idea is now to prove the same limit theorems for the upper and lower bounds using combinatorial techniques.

What amounts to a rescaling argument with a little calculus shows that for a (not very explicit) constant μ,

$$\lim_{n\to\infty} \frac{\mathbb{E}[\widetilde{S}_{n,k+1}]}{n^{k+2} r_n^{d(k+1)}} = \lim_{n\to\infty} \frac{\mathbb{E}[S_{n,k+1} + Y_{n,k+1} + Z_{n,k+1}]}{n^{k+2} r_n^{d(k+1)}} = \frac{\mu}{(k+1)!}.$$

$$\Longrightarrow \qquad \mathbb{P}[\beta_k(\mathcal{C}) \geq 1] \leq \mathbb{E}[\beta_k(\mathcal{C})] \lesssim n^{k+2} r_n^{d(k+1)};$$

$$\widetilde{S}_{n,k+1} \leq \beta_k(\mathcal{C}(X_1, \ldots, X_n)) \leq S_{n,k+1} + Y_{n,k+1} + Z_{n,k+1}$$

The idea is now to prove the same limit theorems for the upper and lower bounds using combinatorial techniques.

What amounts to a rescaling argument with a little calculus shows that for a (not very explicit) constant μ,

$$\lim_{n\to\infty} \frac{\mathbb{E}[\widetilde{S}_{n,k+1}]}{n^{k+2} r_n^{d(k+1)}} = \lim_{n\to\infty} \frac{\mathbb{E}[S_{n,k+1} + Y_{n,k+1} + Z_{n,k+1}]}{n^{k+2} r_n^{d(k+1)}} = \frac{\mu}{(k+1)!}.$$

$$\Longrightarrow \qquad \mathbb{P}[\beta_k(\mathcal{C}) \geq 1] \leq \mathbb{E}[\beta_k(\mathcal{C})] \lesssim n^{k+2} r_n^{d(k+1)};$$

That is, if $n^{k+2} r_n^{d(k+1)} \to 0$, then $\beta_k(\mathcal{C}) \to 0$ a.a.s.

The other end of the sparse regime: $n^{k+2} r_n^{d(k+1)} \to \infty$

The other end of the sparse regime: $n^{k+2} r_n^{d(k+1)} \to \infty$

- In the sparsest case $n^{k+2} r_n^{d(k+1)} \to 0$, there aren't any $(k+1)$-simplices to contribute to β_k.

The other end of the sparse regime: $n^{k+2} r_n^{d(k+1)} \to \infty$

- In the sparsest case $n^{k+2} r_n^{d(k+1)} \to 0$, there aren't any $(k+1)$-simplices to contribute to β_k.

- In the moderate situation $n^{k+2} r_n^{d(k+1)} \to \beta \in (0, \infty)$, there are $(k+1)$-simplices, but analyzing β_k is comparatively straightforward because the $(k+1)$-simplices are essentially always isolated, so we just have to count them.

The other end of the sparse regime: $n^{k+2} r_n^{d(k+1)} \to \infty$

- In the sparsest case $n^{k+2} r_n^{d(k+1)} \to 0$, there aren't any $(k+1)$-simplices to contribute to β_k.

- In the moderate situation $n^{k+2} r_n^{d(k+1)} \to \beta \in (0, \infty)$, there are $(k+1)$-simplices, but analyzing β_k is comparatively straightforward because the $(k+1)$-simplices are essentially always isolated, so we just have to count them.

- Once $n^{k+2} r_n^{d(k+1)} \to \infty$, this is no longer the case; this increased spacial dependence presents significant technical difficulties.

The other end of the sparse regime: $n^{k+2} r_n^{d(k+1)} \longrightarrow \infty$

- In the sparsest case $n^{k+2} r_n^{d(k+1)} \to 0$, there aren't any $(k+1)$-simplices to contribute to β_k.

- In the moderate situation $n^{k+2} r_n^{d(k+1)} \to \beta \in (0, \infty)$, there are $(k+1)$-simplices, but analyzing β_k is comparatively straightforward because the $(k+1)$-simplices are essentially always isolated, so we just have to count them.

- Once $n^{k+2} r_n^{d(k+1)} \to \infty$, this is no longer the case; this increased spacial dependence presents significant technical difficulties.

An important tool in such situations is to "Poissonize" the problem, get the theorem there, and then "de-Poissonize" to get the theorem we're really after.

How to Poissonize your (our) problem

How to Poissonize your (our) problem

Recall that the complex is built starting from n points chosen independently and identically according to a fixed distribution.

How to Poissonize your (our) problem

Recall that the complex is built starting from n points chosen independently and identically according to a fixed distribution.

This produces inherent spacial dependence; i.e., if I tell you what's going on in a particular region, you now know more about what may or may not be happening in other regions.

How to Poissonize your (our) problem

Recall that the complex is built starting from n points chosen independently and identically according to a fixed distribution.

This produces inherent spacial dependence; i.e., if I tell you what's going on in a particular region, you now know more about what may or may not be happening in other regions.

The Poisson process doesn't have this property: a Poisson process with intensity measure μ is a collection of (a random number of) random points in $\mathbb{R}^d$, such that

- the number of points in a region $A \subseteq \mathbb{R}^d$ is a Poisson random variable with mean $\mu(A)$; and
- if A and B are disjoint regions of $\mathbb{R}^d$, then the number of points in A is *independent* of the number of points in B.

We model our i.i.d. points with a Poisson process:

We model our i.i.d. points with a Poisson process:

Let N_n be a Poisson random variable with mean n, and let

$$\{X_1, X_2, \dots\}$$

be a sequence of independent random points, each distributed according to our density f, and independent of N_n. Then

$$\mathcal{P} := \{X_1, \dots, X_{N_n}\}$$

is a Poisson process with intensity $nf(\cdot)$.

The Poissonized problem

We consider the Poissonized random variables

$$\widetilde{S}^P_{n,k} \qquad \text{and} \qquad S^P_{n,k} + Y^P_{n,k} + Z^P_{n,k},$$

defined as before but over the collection $\mathcal{P} = \{X_1, \dots X_{N_n}\}$.

The Poissonized problem

We consider the Poissonized random variables

$$\widetilde{S}^P_{n,k} \qquad \text{and} \qquad S^P_{n,k} + Y^P_{n,k} + Z^P_{n,k},$$

defined as before but over the collection $\mathcal{P} = \{X_1, \ldots X_{N_n}\}$.

Means and variances are harder to compute in this case, but can essentially be recovered from the i.i.d. case.

The Poissonized problem

We consider the Poissonized random variables

$$\widetilde{S}^P_{n,k} \qquad \text{and} \qquad S^P_{n,k} + Y^P_{n,k} + Z^P_{n,k},$$

defined as before but over the collection $\mathcal{P} = \{X_1, \dots X_{N_n}\}$.

Means and variances are harder to compute in this case, but can essentially be recovered from the i.i.d. case.

The matching upper and lower normal approximation theorems are proved via the dependency graph approach to Stein's method – for more on this story, stay tuned until next time.