Limit Theorems for Betti Numbers of Extreme Sample Clouds

Takashi Owada

Technion-Israel Institute of Technology

June 2016

Topological data analysis

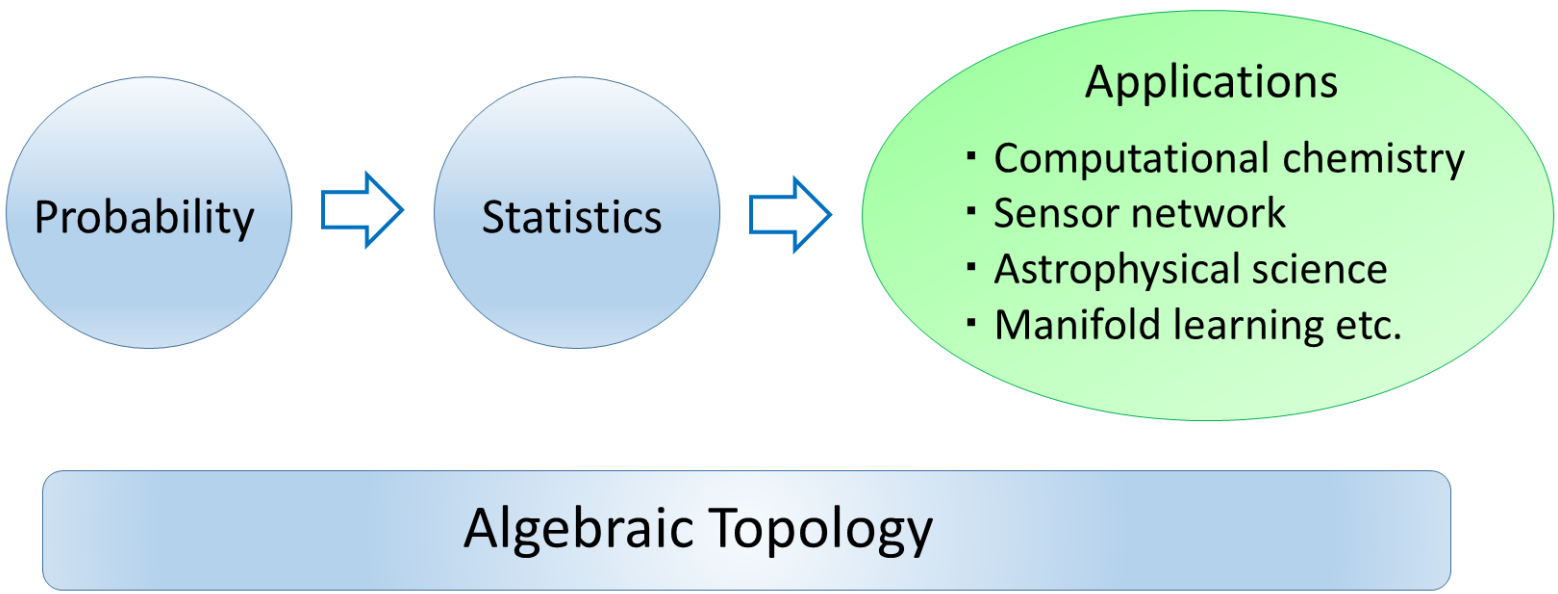
- Topological data analysis (TDA) is an approach to the analysis of datasets using techniques from topology and other mathematics.

- Typically, topologists classify objects into classes of "similar shapes" by the number of holes.

Topological data analysis

- Topological data analysis (TDA) is an approach to the analysis of datasets using techniques from topology and other mathematics.

- Typically, topologists classify objects into classes of "similar shapes" by the number of holes.



- As highlighted in a recent series of columns in the IMS Bulletin, the collaboration of three different disciplines, topology, probability, and statistics, is indispensable for the development of TDA.
  - The author of the column has invented a word, TOPOS (=topology, probability, and statistics).
- However, there are still only limited number of probabilistic and statistical works in TDA.

Betti numbers

- Basic quantifier in algebraic topology.
- Given a topological space X, the 0-th Betti number $\beta_0(X)$ is defined as

$$\beta_0(X) = \text{the number of connected components in } X.$$

- For $k \geq 1$, the k-th Betti number $\beta_k(X)$ is defined as

$$\beta_k(X) = \text{the number of } k\text{-dim holes in } X.$$

Betti numbers

- Basic quantifier in algebraic topology.
- Given a topological space X, the 0-th Betti number $\beta_0(X)$ is defined as

$$\beta_0(X) = \text{the number of connected components in } X.$$

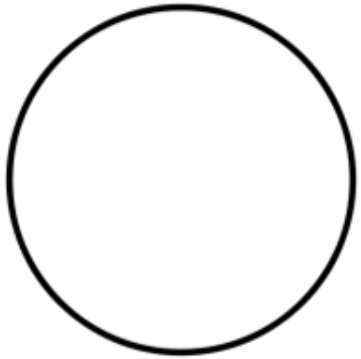
- For $k \geq 1$, the k-th Betti number $\beta_k(X)$ is defined as

$$\beta_k(X) = \text{the number of } k\text{-dim holes in } X.$$

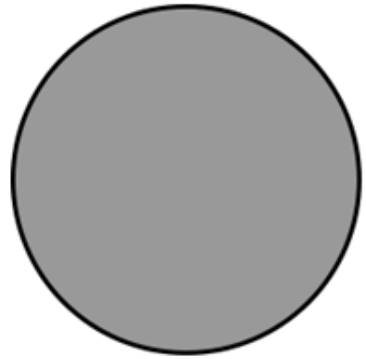
  - More intuitively,

$$\beta_1(X) = \text{the number of ``closed loops'' in } X.$$

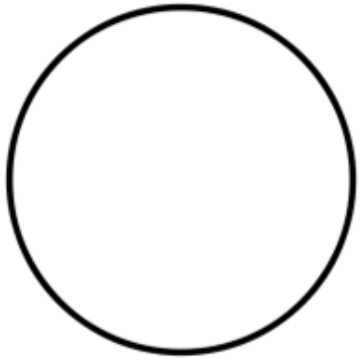
$$\beta_2(X) = \text{the number of ``hollows'' in } X.$$



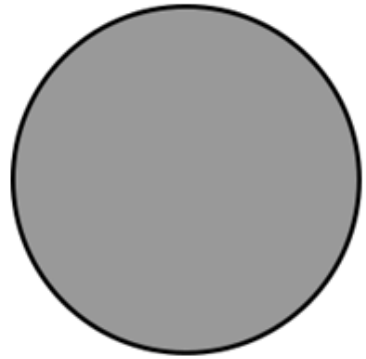
$\beta_0 = 1,\ \beta_1 = 1,$
$\beta_k = 0,\ k \geq 2.$



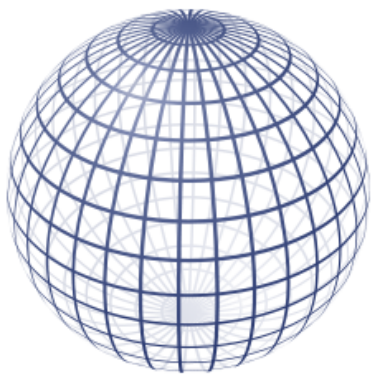
$\beta_0 = 1,\ \beta_k = 0,\ k \geq 1.$



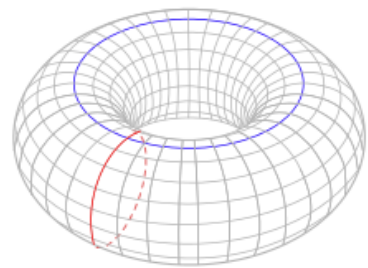
$\beta_0 = 1,\ \beta_1 = 1,$
$\beta_k = 0,\ k \geq 2.$



$\beta_0 = 1,\ \beta_k = 0,\ k \geq 1.$



$\beta_0 = 1,\ \beta_1 = 0,\ \beta_2 = 1,$
$\beta_k = 0,\ k \geq 3.$



$\beta_0 = 1,\ \beta_1 = 2,\ \beta_2 = 1,$
$\beta_k = 0,\ k \geq 3.$

Scheme

1. Generate random sample from an underlying probability law.

   - (X_i): iid $\mathbb{R}^d$-valued random variables, $d \geq 2$, with spherically symmetric density f.

Scheme

1. Generate random sample from an underlying probability law.

- (X_i): iid $\mathbb{R}^d$-valued random variables, $d \geq 2$, with spherically symmetric density f.

Case I: regularly varying tail

- f has a regularly varying tail: for some $\alpha > d$,

$$\lim_{r \to \infty} \frac{f(rte_1)}{f(re_1)} = t^{-\alpha} \quad \text{for every } t > 0\,,$$

where $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^d$.

Scheme

1. Generate random sample from an underlying probability law.

- (X_i): iid $\mathbb{R}^d$-valued random variables, $d \geq 2$, with spherically symmetric density f.

Case I: regularly varying tail

- f has a regularly varying tail: for some $\alpha > d$,

$$\lim_{r\to\infty} \frac{f(rte_1)}{f(re_1)} = t^{-\alpha} \quad \text{for every } t > 0\,,$$

  where $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^d$.

- For example,

$$f(x) = C/\big(1 + ||x||^\alpha\big)\,, \quad x \in \mathbb{R}^d,\ \alpha > d\,.$$

Case II: exponentially decaying tail

- Assume that $f(x) = L\big(\|x\|\big)e^{-\psi(\|x\|)}$, $x \in \mathbb{R}^d$.
  - $\psi'(z) > 0$, $\psi(z) \to \infty$, $(1/\psi')'(z) \to 0$ as $z \to \infty$.
  - ψ is a von-Mises function.
  - $L : \mathbb{R}_+ \to \mathbb{R}_+$ can be ignored in the tail of f.

Case II: exponentially decaying tail

- Assume that $f(x) = L\big(\|x\|\big)e^{-\psi(||x||)}$, $x \in \mathbb{R}^d$.
  - $\psi'(z) > 0$, $\psi(z) \to \infty$, $(1/\psi')'(z) \to 0$ as $z \to \infty$.
  - ψ is a von-Mises function.
  - $L : \mathbb{R}_+ \to \mathbb{R}_+$ can be ignored in the tail of f.
- For example,

$$f(x) = Ce^{-||x||^\tau/\tau}, \quad x \in \mathbb{R}^d, \quad \tau > 0.$$

To make our story simpler, we will work on special examples in the following.

Case I:

$$f(x) = C/\left(1 + ||x||^{\alpha}\right), \quad x \in \mathbb{R}^d, \ \alpha > d\,.$$

To make our story simpler, we will work on special examples in the following.

Case I:

$$f(x) = C/\left(1 + ||x||^{\alpha}\right), \quad x \in \mathbb{R}^d, \ \alpha > d\,.$$

Case II:

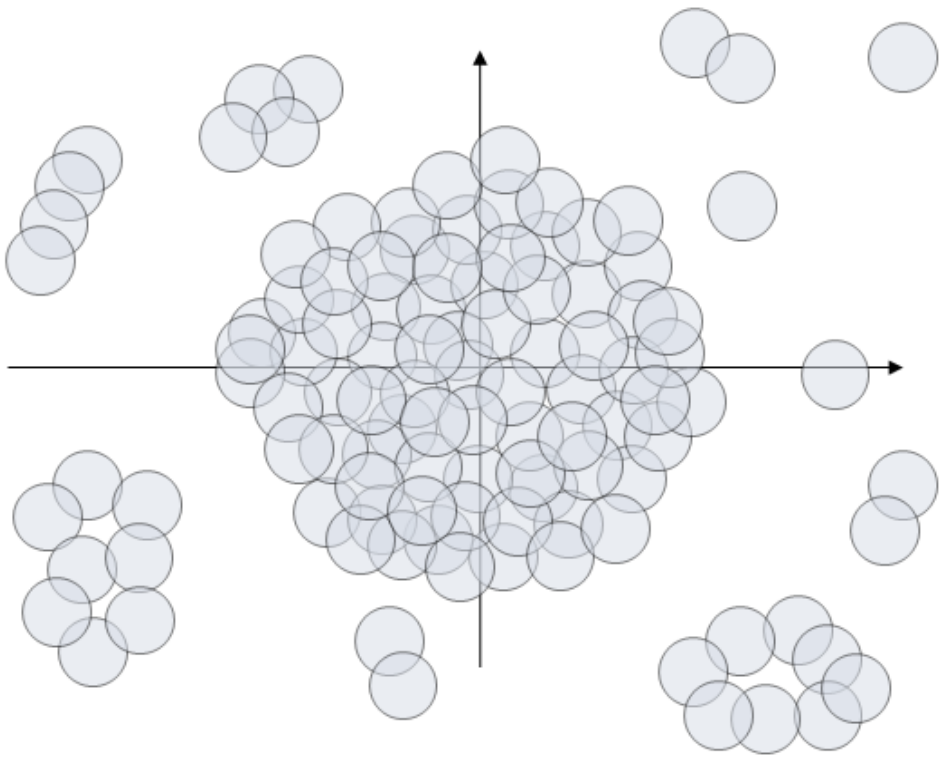
$$f(x) = Ce^{-||x||^{\tau}/\tau}, \quad x \in \mathbb{R}^d, \quad \tau > 0.$$

- If $0 < \tau < 1$, f has a subexponential tail.
- If $\tau = 1$, f has an exponential tail.
- If $\tau > 1$, f has a superexponential tail.

Note: All the results in this talk hold in the general setup.

2. Draw random balls of radius t about X_i's.

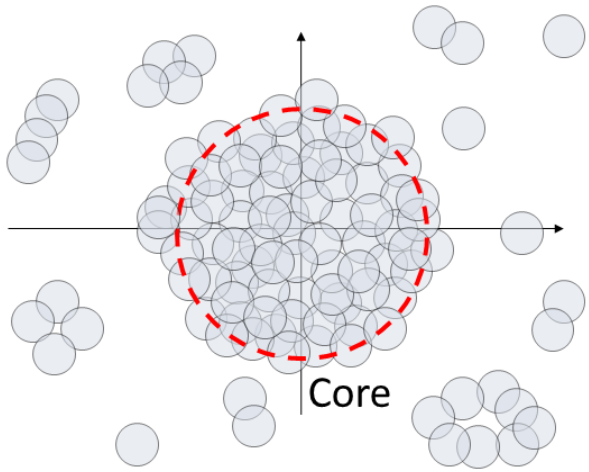
3. Establish the limit theorems for Betti numbers of the union of balls.

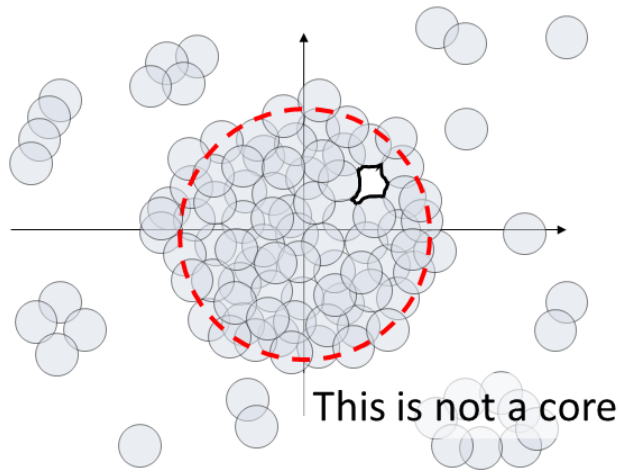


$\beta_0(\text{union of balls}) = 11$

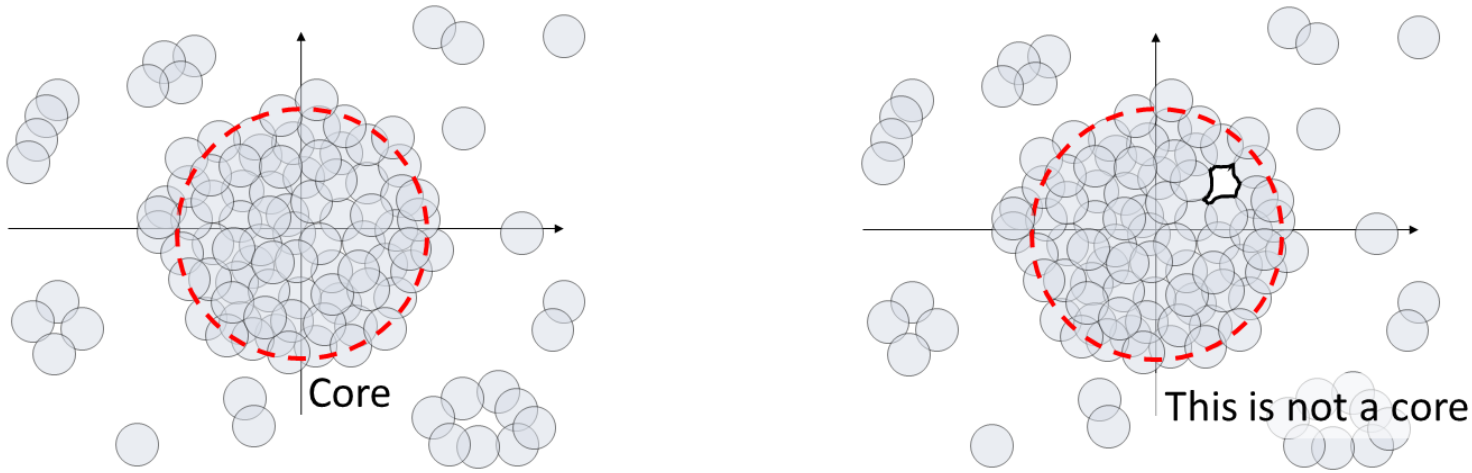
$\beta_1(\text{union of balls}) = 3$

- Core = a centered ball in which random points are scattered very densely so that the union of balls inside the region is contractible (= can continuously shrink to a single point).





- Core = a centered ball in which random points are scattered very densely so that the union of balls inside the region is contractible (= can continuously shrink to a single point).



Formal definition.

Given a set of random points $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, a centered ball $B(0; R_n)$ is called a core if

$$\mathbb{P}\left\{ B(0;R_n) \subset \bigcup_{X \in \mathcal{X}_n \cap B(0;R_n)} B(X;1) \right\} \to 1, \quad n \to \infty.$$

Proposition [Adler et. al, 2014].

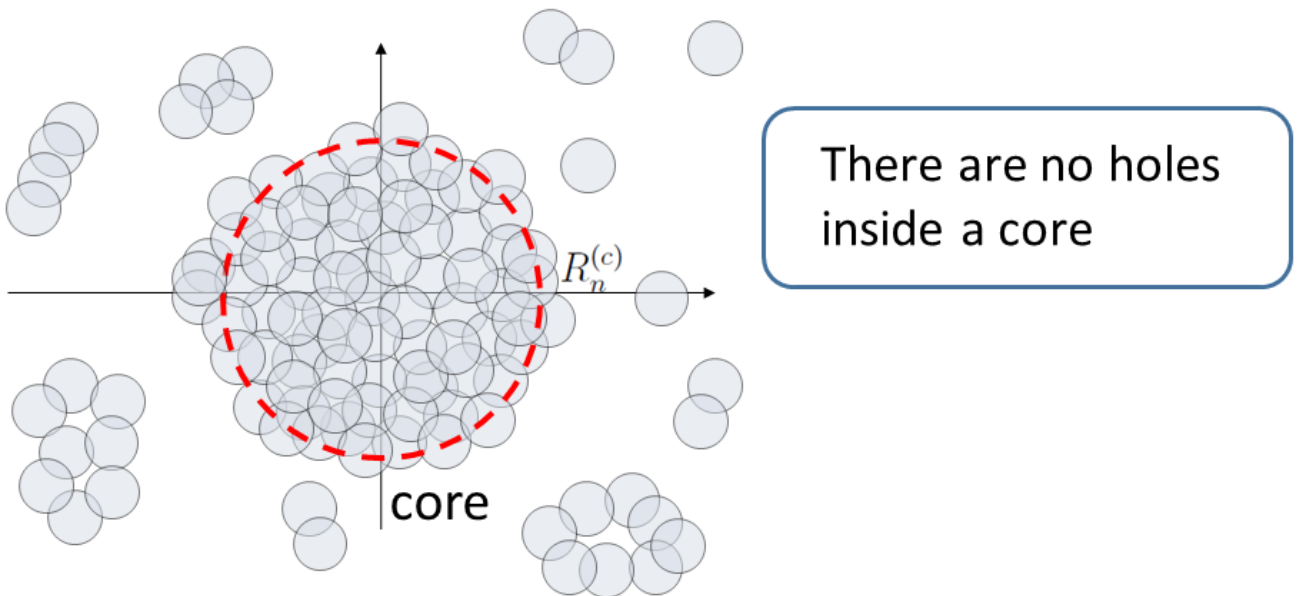
Let $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ be an iid random sample with density

$$f(x) = C/\left(1 + ||x||^{\alpha}\right), \quad x \in \mathbb{R}^d, \ \alpha > d\,.$$

Then, there exists a sequence $R_n^{(c)} = C'(n/\log n)^{1/\alpha}$ for some $C' > 0$, such that

$$\mathbb{P}\left\{ B(0; R_n^{(c)}) \subset \bigcup_{X \in \mathcal{X}_n \cap B(0; R_n^{(c)})} B(X; 1) \right\} \to 1\,, \quad n \to \infty,$$

i.e., $B(0; R_n^{(c)})$ is a core.



Proposition [Adler et. al, 2014].

Let $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ be an iid random sample with density

$$f(x) = Ce^{-||x||^\tau/\tau}, \quad x \in \mathbb{R}^d, \quad \tau > 0.$$

Then,

$$\mathbb{P}\left\{ B(0; R_n^{(c)}) \subset \bigcup_{X \in \mathcal{X}_n \cap B(0; R_n^{(c)})} B(X; 1) \right\} \to 1\,, \quad n \to \infty,$$

holds, where

$$R_n^{(c)} = \left(\tau \log n - \tau \log\log(\tau \log n)^{1/\tau} + C'\right)^{1/\tau}.$$

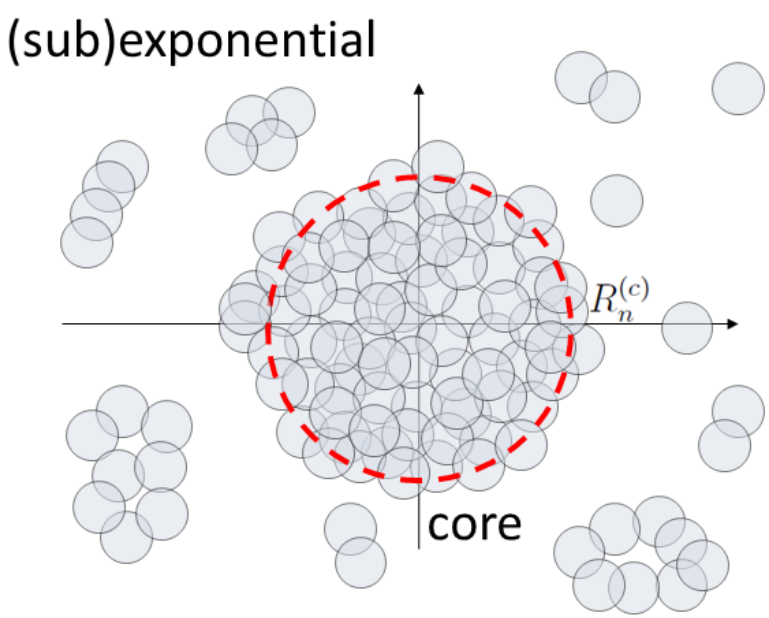
Proposition [Adler et. al, 2014] (continued).

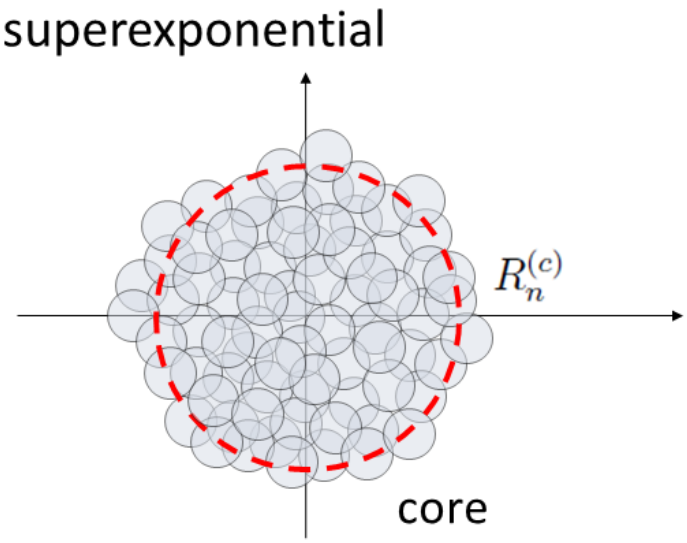
(i) If $0 < \tau \leq 1$, i.e, f has either a subexponential or an exponential tail,

$$\mathbb{P}\Big\{ \text{there exist random points outside } B(0; R_n^{(c)}) \Big\} \to 1, \quad n \to \infty.$$

(ii) If $\tau > 1$, i.e, f has a superexponential tail,

$$\mathbb{P}\Big\{ \text{there exist random points outside } B(0; R_n^{(c)}) \Big\} \to 0, \quad n \to \infty.$$





- The related notion, a weak core, plays a more decisive role in the characterization of the limit theorems for Betti numbers.

Definition

Let f be a spherically symmetric density on $\mathbb{R}^d$ and $R_n^{(w)} \to \infty$ be a sequence determined by

$$nf\left(R_n^{(w)} e_1\right) \to 1\,, \quad n \to \infty\,.$$

Then $B(0; R_n^{(w)})$ is called a weak core.

Case I (power-law tail)

- If $f(x) = C/\big(1 + \|x\|^{\alpha}\big)$, $x \in \mathbb{R}^d$, $\alpha > d$, then

$$R_n^{(w)} = (Cn)^{1/\alpha}.$$

  c.f. $R_n^{(c)} = C'\big(n/\log n\big)^{1/\alpha}$.

- $R_n^{(c)}/R_n^{(w)} \to 0$, but they share the same regular variation exponent, $1/\alpha$.

Case I (power-law tail)

- If $f(x) = C/\big(1 + \|x\|^{\alpha}\big)$, $x \in \mathbb{R}^d$, $\alpha > d$, then

$$R_n^{(w)} = (Cn)^{1/\alpha}.$$

c.f. $R_n^{(c)} = C'\big(n/\log n\big)^{1/\alpha}$.

- $R_n^{(c)}/R_n^{(w)} \to 0$, but they share the same regular variation exponent, $1/\alpha$.

Case II (exponentially decaying tail)

- If $f(x) = Ce^{-\|x\|^{\tau}/\tau}$, $x \in \mathbb{R}^d$, $\tau > 0$, then

$$R_n^{(w)} = (\tau \log n + \tau \log C)^{1/\tau}.$$

c.f. $R_n^{(c)} = \big(\tau \log n - \tau \log\log(\tau \log n)^{1/\tau} + C'\big)^{1/\tau}$.

Betti number in the tail

- $\mathcal{X}_n = \{X_1, \ldots, X_n\}$: iid $\mathbb{R}^d$-valued random variables drawn from a power-law distribution or a (sub)exponential distribution.
- For $k \geq 1$, define

$$\beta_{k,n}(t) := \beta_k\left(\bigcup_{X \in \mathcal{X}_n \setminus B(0;R_n)} B(X;t)\right), \quad t \geq 0\,,$$

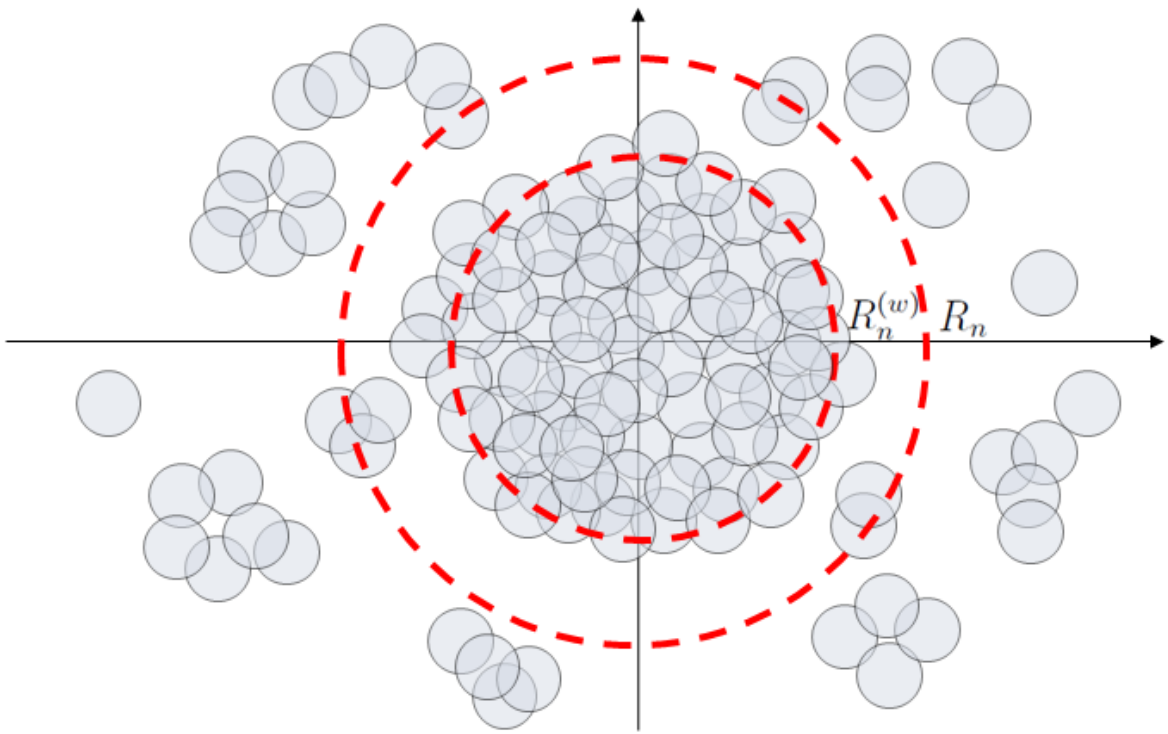
where R_n is a non-random sequence with $R_n \geq R_n^{(w)}$ (= radius of a weak core).

Betti number in the tail

- $\mathcal{X}_n = \{X_1, \ldots, X_n\}$: iid $\mathbb{R}^d$-valued random variables drawn from a power-law distribution or a (sub)exponential distribution.
- For $k \geq 1$, define

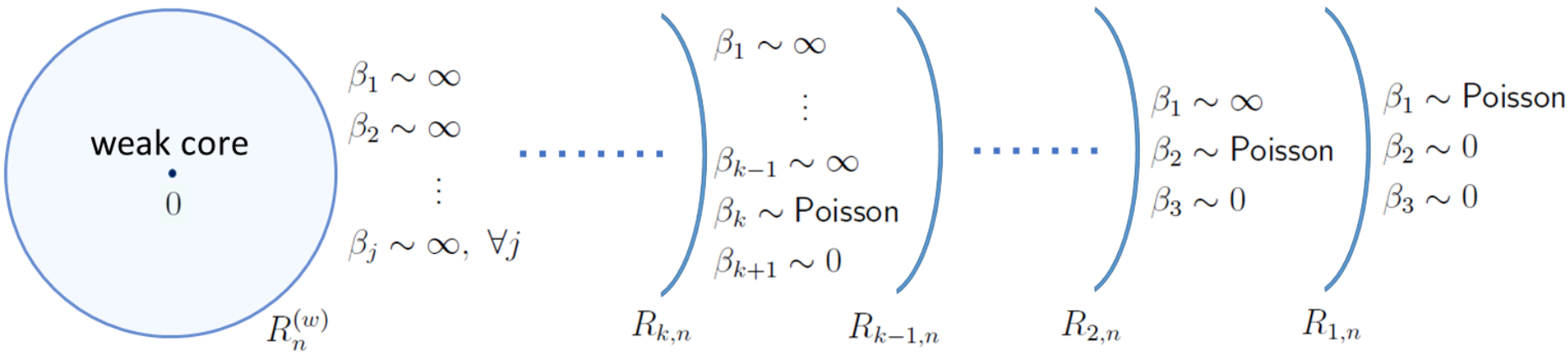
$$\beta_{k,n}(t) := \beta_k\left(\bigcup_{X \in \mathcal{X}_n \setminus B(0;R_n)} B(X;t)\right), \quad t \geq 0\,,$$

where R_n is a non-random sequence with $R_n \geq R_n^{(w)}$ (= radius of a weak core).



Layered non-trivial homologies

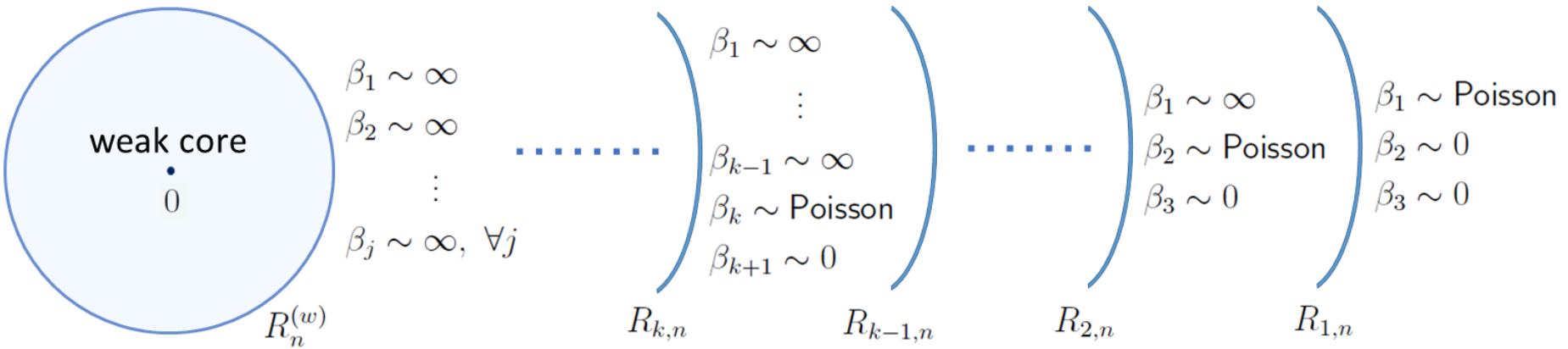
- Assume $f(x) = C/\big(1 + \|x\|^{\alpha}\big)$, $x \in \mathbb{R}^d$, $\alpha > d$.
- Asymptotically (i.e., $n \to \infty$), we observe the following layered structure.



$$R_n^{(w)} = (Cn)^{1/\alpha}$$
$$R_{k,n} = (Cn)^{1/(\alpha - d/(k+2))}$$

Layered non-trivial homologies

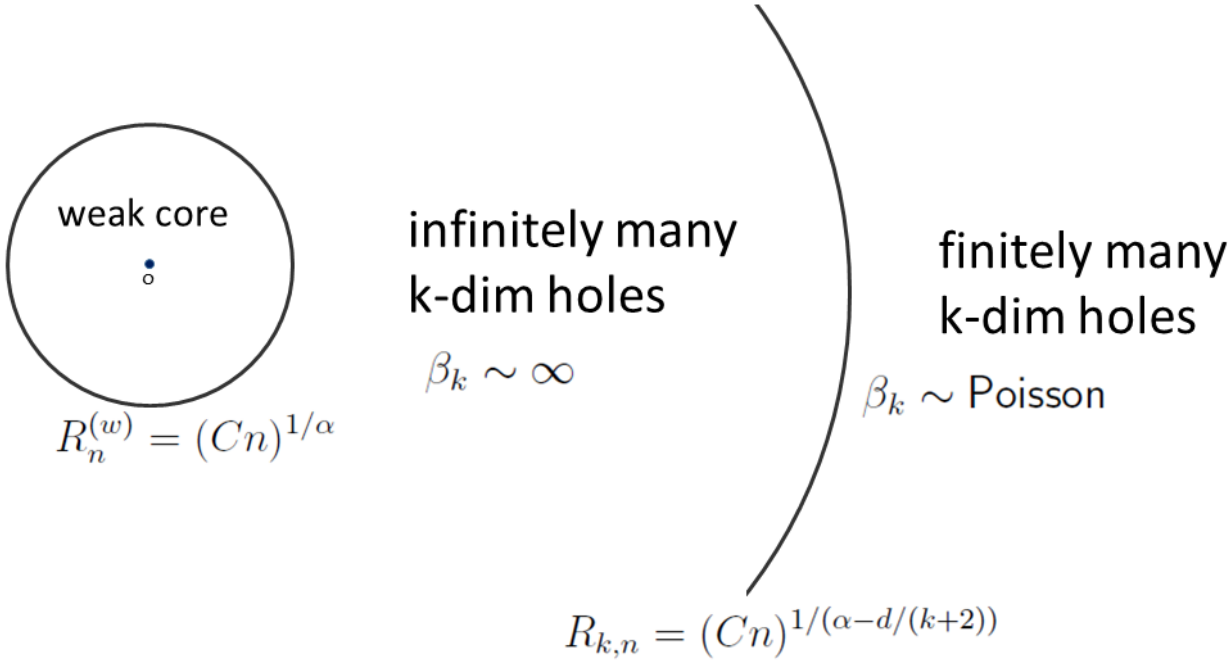
- Assume $f(x) = Ce^{-\|x\|^\tau/\tau}$, $x \in \mathbb{R}^d$, $0 < \tau \leq 1$ (i.e., having a (sub)exponential tail).
- Asymptotically (i.e., $n \to \infty$), we observe the following layered structure.



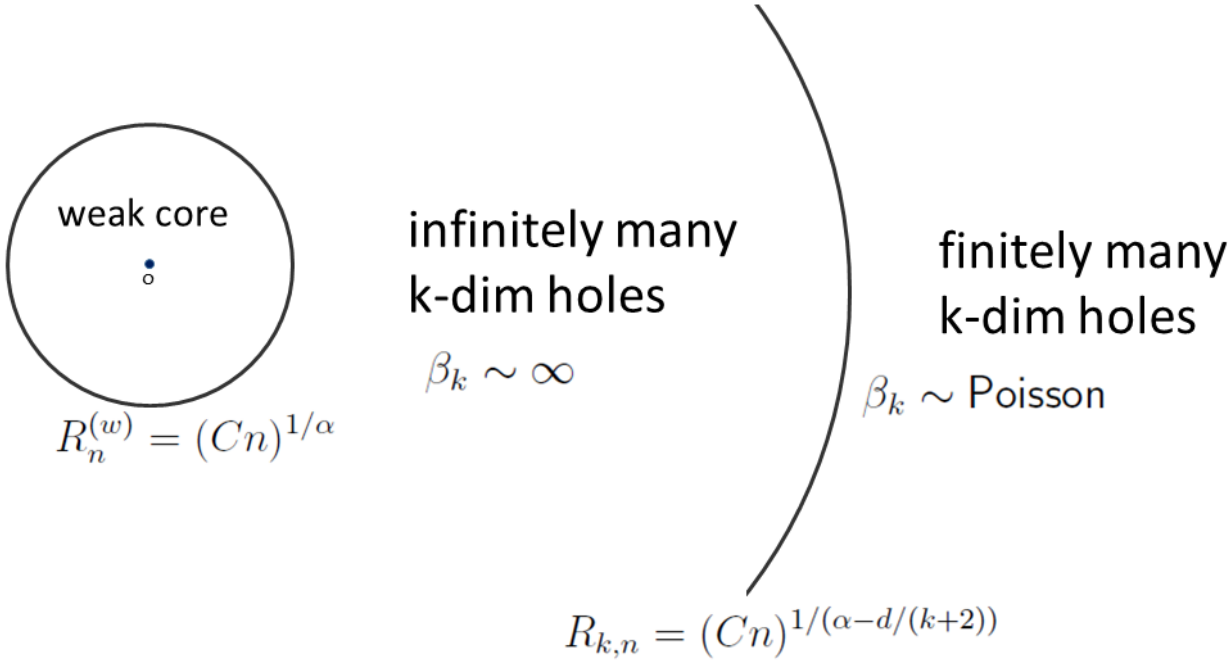
$$R_n^{(w)} = (\tau \log n + \tau \log C)^{1/\tau}$$
$$R_{k,n} = \left(\tau \log n + (k+2)^{-1}(d-\tau)\log(\tau \log n) + \tau \log C\right)^{1/\tau}$$

- In what follows, we only treat power-law tail distributions.

- Looking at the asymptotic (i.e., $n \to \infty$) spatial distribution of k-dim holes,



- Looking at the asymptotic (i.e., $n \to \infty$) spatial distribution of k-dim holes,

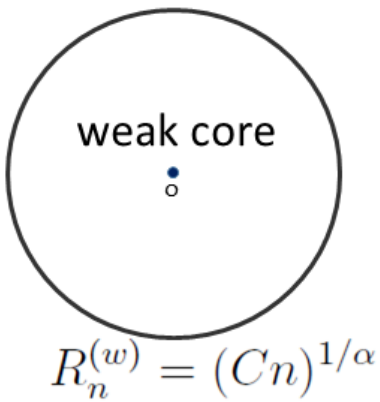


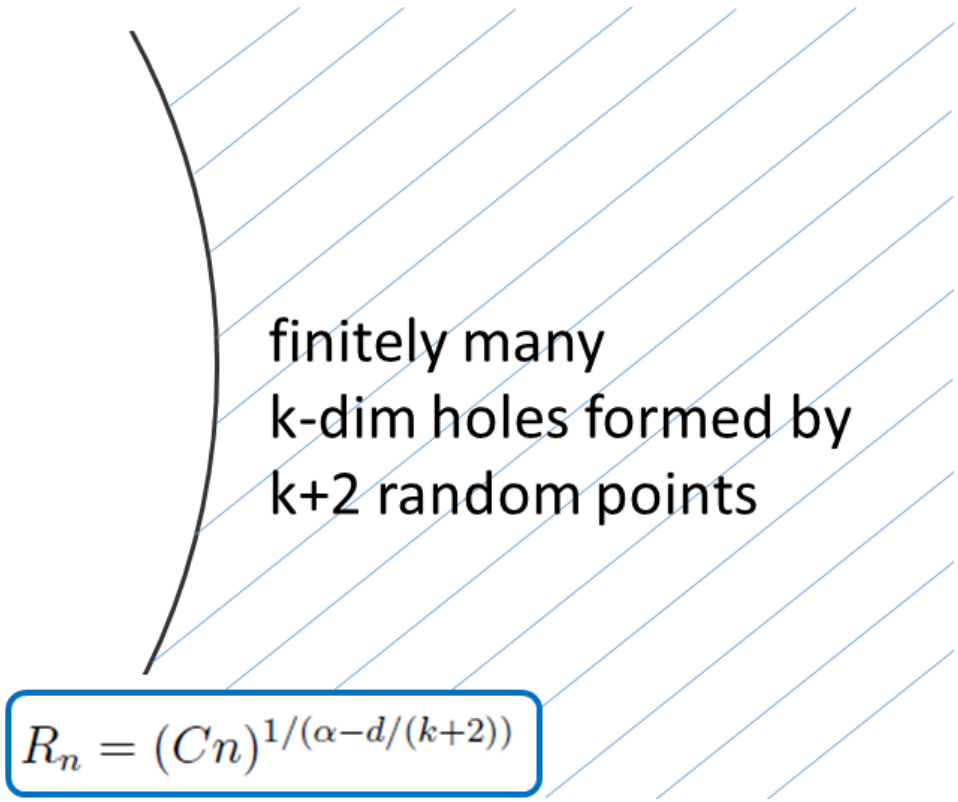
- Three different regimes must be considered.
  We set, respectively,
  - [1]: $R_n = (Cn)^{1/(\alpha - d/(k+2))}$,
  - [2]: $(Cn)^{1/\alpha} \ll R_n \ll (Cn)^{1/(\alpha - d/(k+2))}$,
  - [3]: $R_n = (Cn)^{1/\alpha}$,

  and compute $\beta_{k,n}(t)$ by counting k-dim holes outside $B(0; R_n)$.

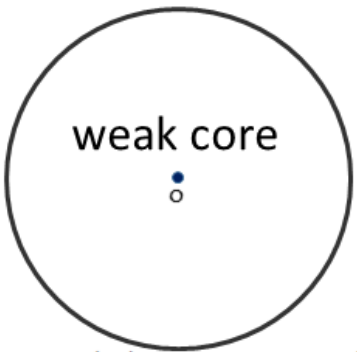
In the regime $[1]$,

- There exist finitely many k-dim holes formed by $k+2$ random points outside $B(0; R_n)$ as $n \to \infty$.

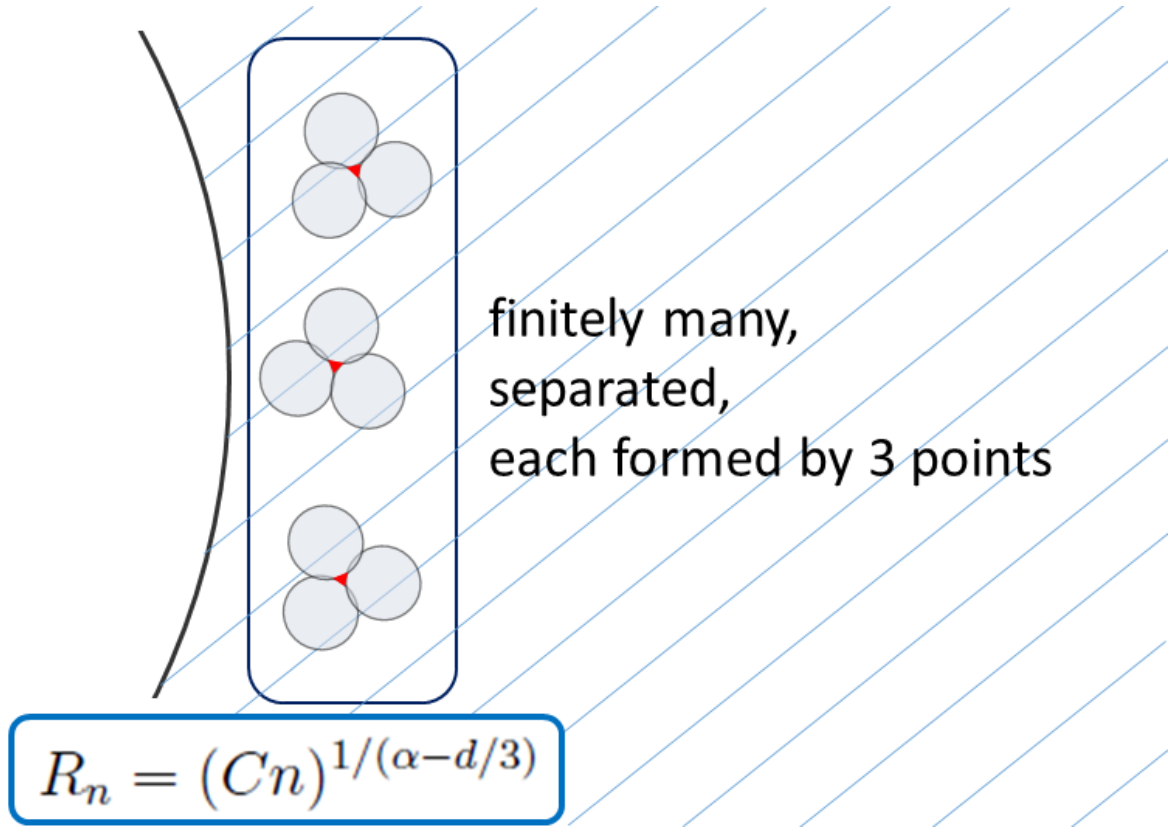




Example $(k = 1)$



$R_n^{(w)} = (Cn)^{1/\alpha}$



$R_n = (Cn)^{1/(\alpha - d/3)}$

- The appearance of holes is a rare event.

Limiting process for $\beta_{k,n}(t)$:

$$N_k(t) := \int_{(\mathbb{R}^d)^{k+1}} h_t(0, y_1, \ldots, y_{k+1})\, M_k(d\mathbf{y}).$$

- M_k is a Poisson random measure with Lebesgue intensity measure on $(\mathbb{R}^d)^{k+1}$.
- $h_t(0, y_1, \ldots, y_{k+1}) = \mathbf{1}\left\{ \beta_k\left(B(0;t) \cup \bigcup_{i=1}^{k+1} B(y_i;t) \right) = 1 \right\}$

  with $0, y_1, \ldots, y_{k+1} \in \mathbb{R}^d$.

- $h_t(0, \mathbf{y})$ can be expressed as

$$h_t(0, \mathbf{y}) = h_t^+(0, \mathbf{y}) - h_t^-(0, \mathbf{y}),$$

where h_t^+ and h_t^- are some other indicator functions, which are non-decreasing in t.

- $h_t(0, \mathbf{y})$ can be expressed as

$$h_t(0, \mathbf{y}) = h_t^+(0, \mathbf{y}) - h_t^-(0, \mathbf{y}),$$

where h_t^+ and h_t^- are some other indicator functions, which are non-decreasing in t.

- Accordingly,

$$\begin{aligned} N_k(t) &= \int_{(\mathbb{R}^d)^{k+1}} h_t^+(0, \mathbf{y})\, M_k(d\mathbf{y}) - \int_{(\mathbb{R}^d)^{k+1}} h_t^-(0, \mathbf{y})\, M_k(d\mathbf{y}) \\ &:= N_k^+(t) - N_k^-(t). \end{aligned}$$

- We can prove that $N_k^+(t)$ and $N_k^-(t)$ are represented as a (time-changed) Poisson process.
  - However, $N_k(t)$ is not a (time-changed) Poisson process.

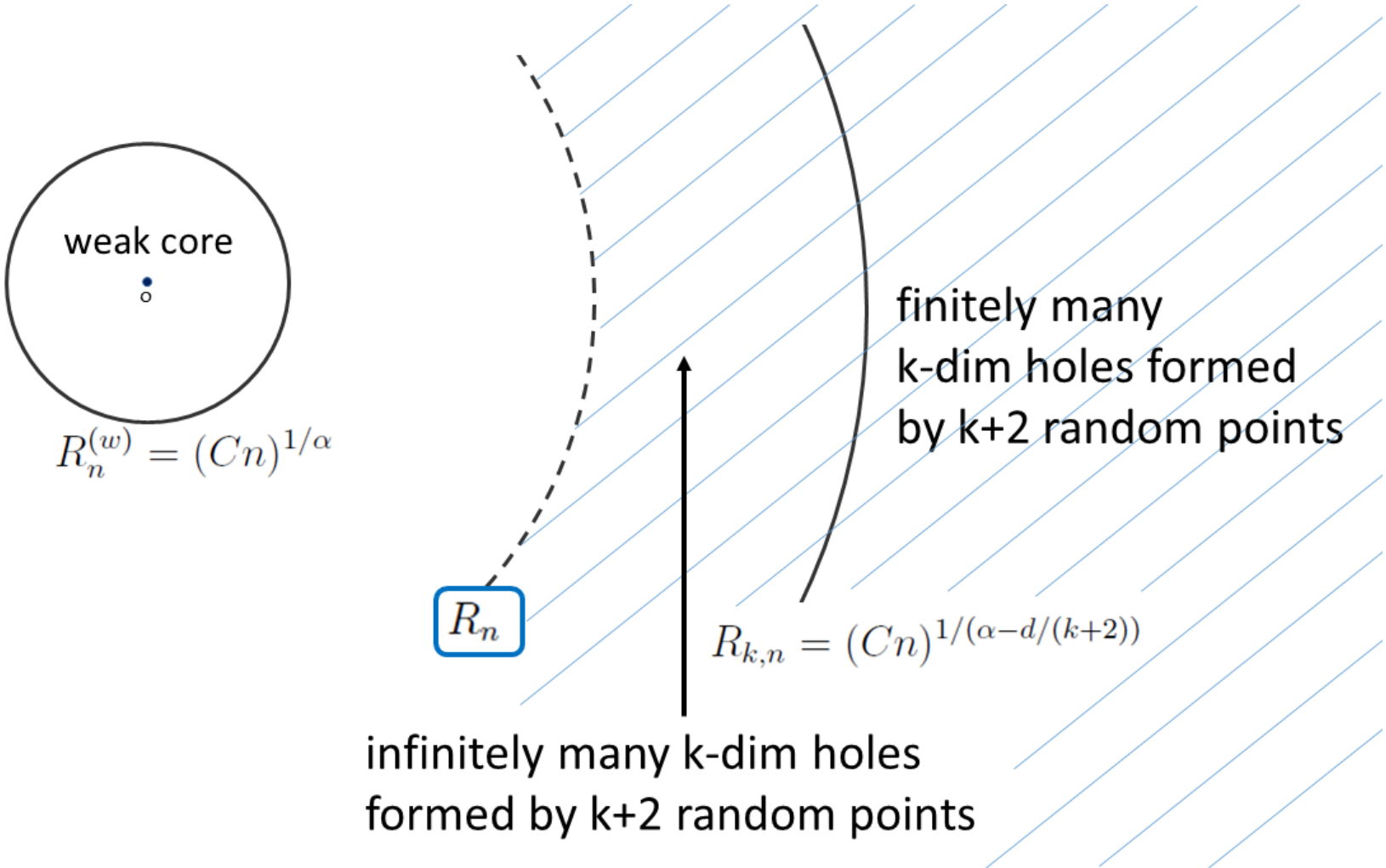
Theorem 1. [O., 2016]

In the regime $[1]$, we have, as $n \to \infty$,

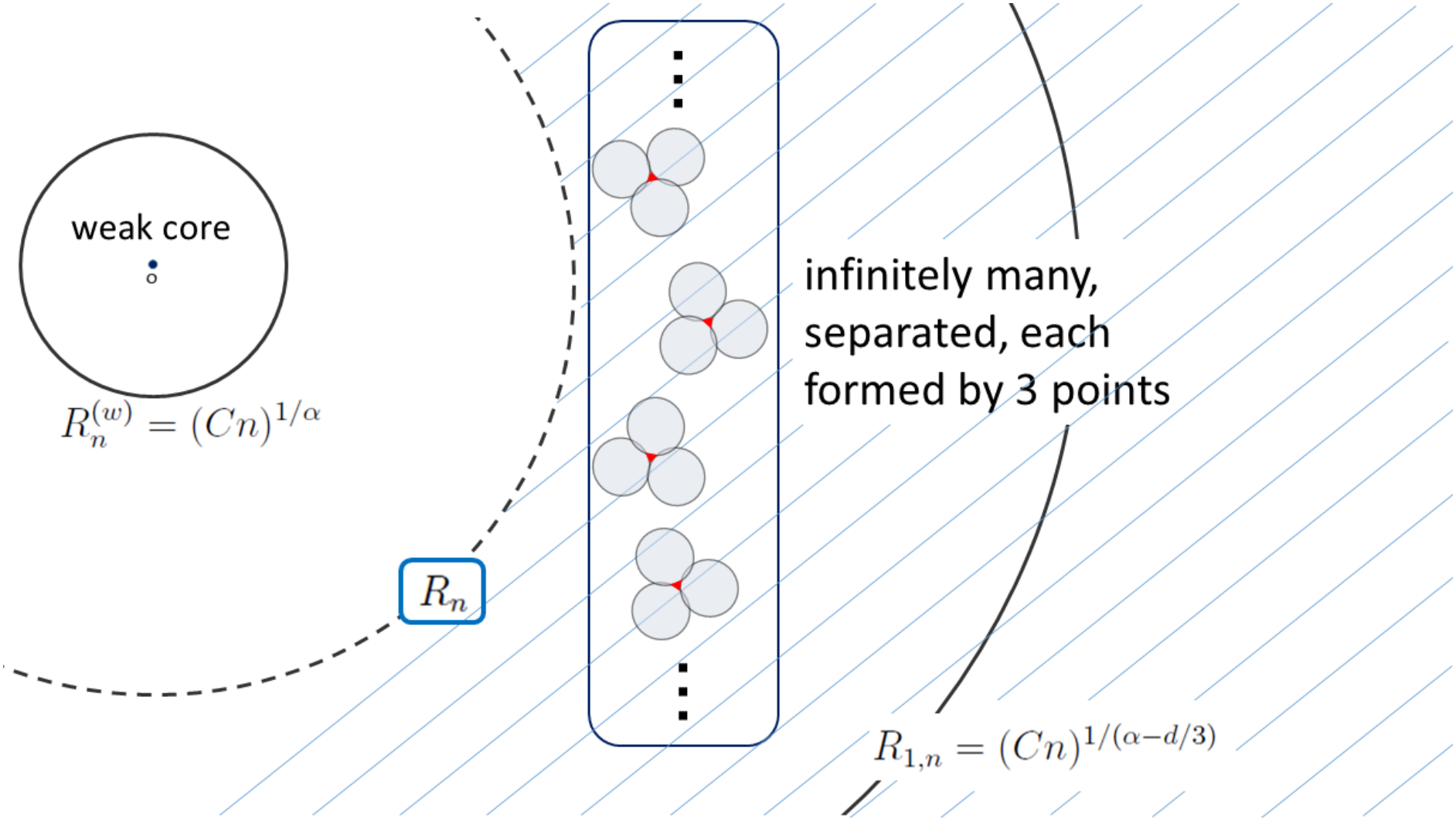
$$\beta_{k,n}(t) \Rightarrow N_k^+(t) - N_k^-(t).$$

In the regime [2],

- There exist infinitely many k-dim holes formed by $k+2$ points outside $B(0; R_n)$ as $n \to \infty$.



Example $(k = 1)$



- The appearance of holes is no longer a rare event.

Limiting process for $\beta_{k,n}(t)$: Define a Gaussian process

$$Y_k(t) := \int_{(\mathbb{R}^d)^{k+1}} h_t(0, y_1, \ldots, y_{k+1})\, G_k(d\mathbf{y}).$$

- G_k is a Gaussian random measure with Lebesgue control measure on $(\mathbb{R}^d)^{k+1}$.
- $h_t(0, y_1, \ldots, y_{k+1}) = \mathbf{1}\left\{ \beta_k\left(B(0;t) \cup \bigcup_{i=1}^{k+1} B(y_i;t) \right) = 1 \right\}$

  with $0, y_1, \ldots, y_{k+1} \in \mathbb{R}^d$.

- Using the decomposition $h_t = h_t^+ - h_t^-$, we can write

$$Y_k(t) = \int_{(\mathbb{R}^d)^{k+1}} h_t^+(0, \mathbf{y})\, G_k(d\mathbf{y}) - \int_{(\mathbb{R}^d)^{k+1}} h_t^-(0, \mathbf{y})\, G_k(d\mathbf{y})$$
$$:= Y_k^+(t) - Y_k^-(t).$$

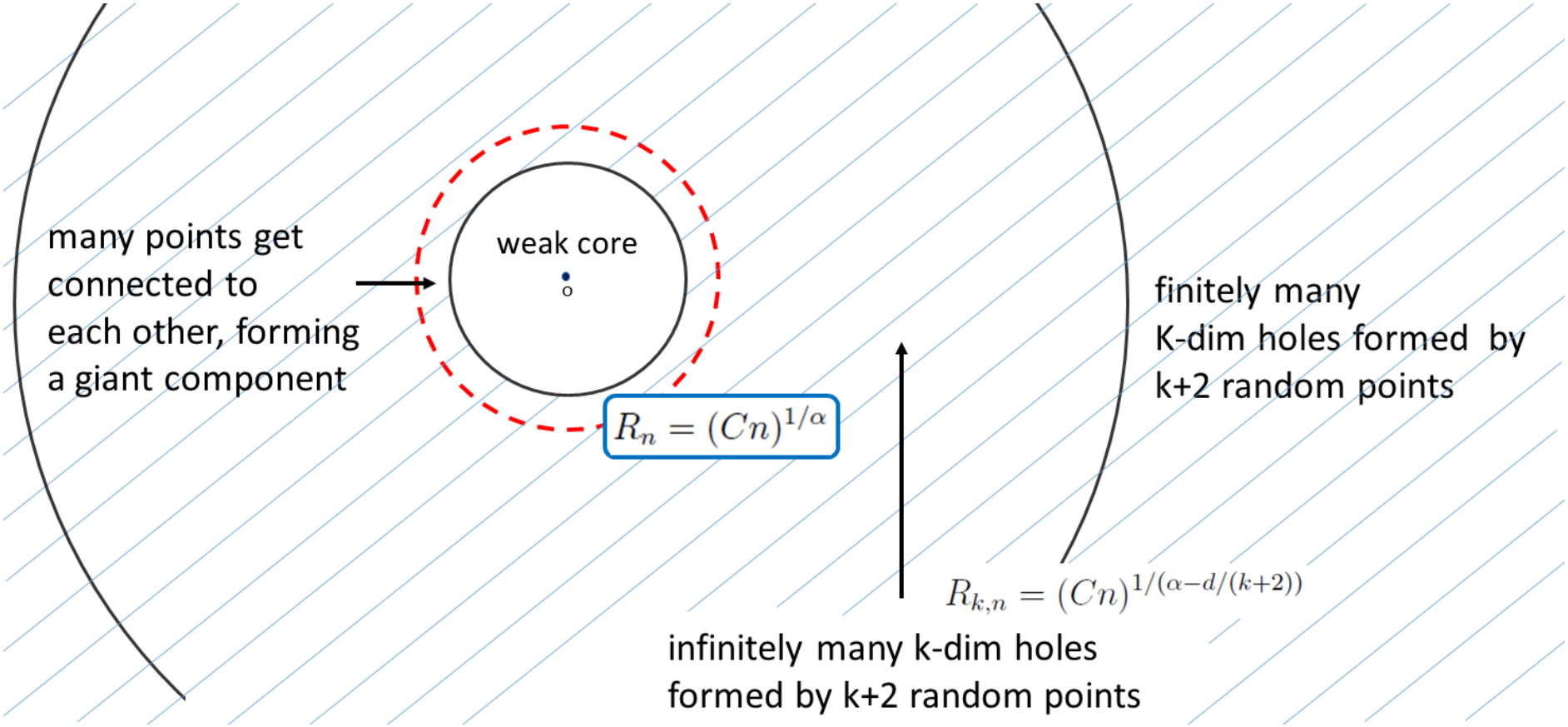
- Then, $Y_k^+(t)$ and $Y_k^-(t)$ are represented as a (time-changed) Brownian motion.
  - $Y_k(t)$ is a Gaussian process, but it is not a (time-changed) Brownian motion.

Theorem 2. [O., 2016]

In the regime $[2]$, we have, as $n \to \infty$,

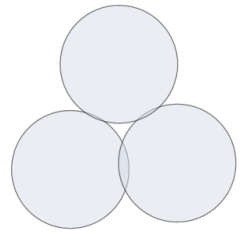
$$\frac{\beta_{k,n}(t) - \mathbb{E}\{\beta_{k,n}(t)\}}{\left(n^{k+2} R_n^{d-\alpha(k+2)}\right)^{1/2}} \Rightarrow Y_k^+(t) - Y_k^-(t).$$

In the regime $[3]$,

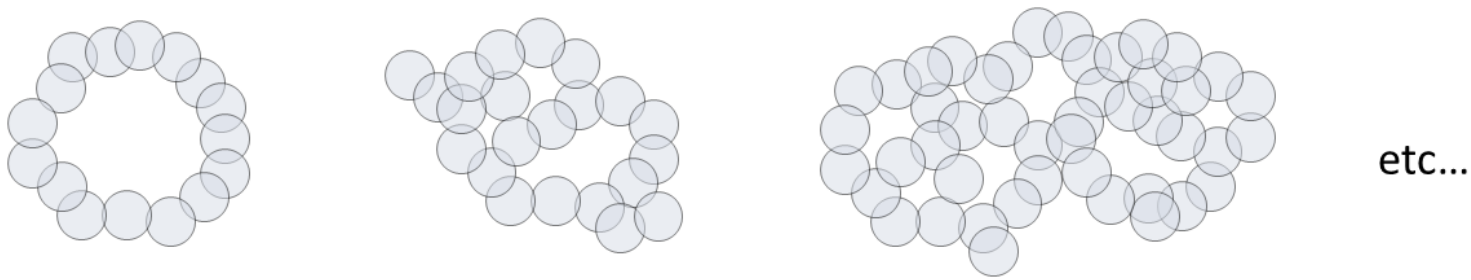


Example $(k = 1)$

- In the regimes $[1]$ and $[2]$, all the one-dim holes contributing to $\beta_{1,n}(t)$ in the limit are always of the form



- In the regime $[3]$, many different kinds of one-dim holes (which exist close enough to a weak core) contribute to $\beta_{1,n}(t)$ in the limit.



The limiting Gaussian process is given by

$$Z_k(t) := \sum_{i=k+2}^{\infty} \sum_{j>0} Z_k^{(i,j)}(t)\,.$$

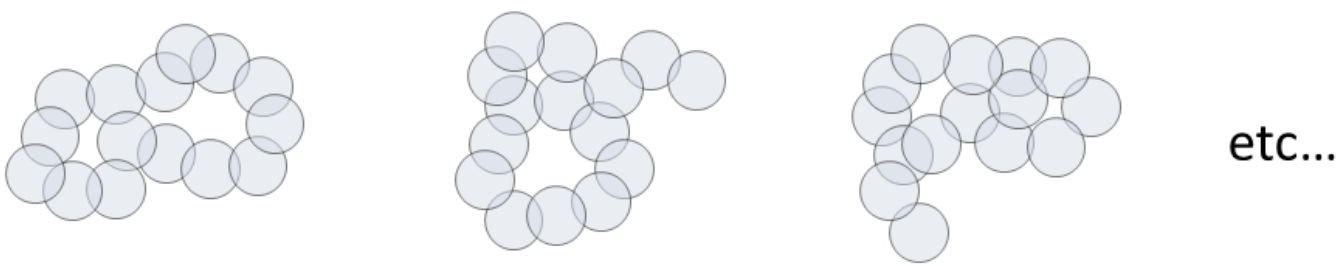
- $Z_k^{(i,j)}(t)$ is a Gaussian process representing the connected components that are formed by i points and contain j holes.

The limiting Gaussian process is given by

$$Z_k(t) := \sum_{i=k+2}^{\infty} \sum_{j>0} Z_k^{(i,j)}(t)\,.$$

- $Z_k^{(i,j)}(t)$ is a Gaussian process representing the connected components that are formed by i points and contain j holes.

Example: $Z_1^{(15,2)}(t)$ (i.e., $k = 1$, $i = 15$, $j = 2$).

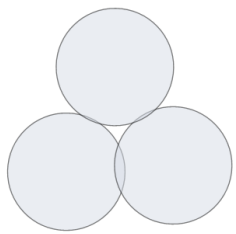


Rewrite $Z_k(t)$ as

$$Z_k(t) = Z_k^{(k+2,1)}(t) + \sum_{i=k+3}^{\infty} \sum_{j>0} Z_k^{(i,j)}(t)\,.$$

- $Z_k^{(k+2,1)}(t)$ represents the connected components that are formed by $k+2$ points and contain a single k-dimensional hole.

Example: $Z_1^{(3,1)}(t)$ (i.e., $k=1$, $i=3$, $j=1$).

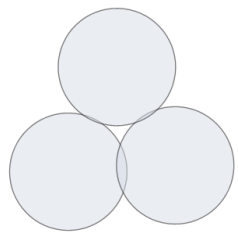


Rewrite $Z_k(t)$ as

$$Z_k(t) = Z_k^{(k+2,1)}(t) + \sum_{i=k+3}^{\infty} \sum_{j>0} Z_k^{(i,j)}(t)\,.$$

- $Z_k^{(k+2,1)}(t)$ represents the connected components that are formed by $k+2$ points and contain a single k-dimensional hole.

Example: $Z_1^{(3,1)}(t)$ (i.e., $k=1$, $i=3$, $j=1$).



- $Z_k^{(k+2,1)}(t)$ is "similar" to the $Y_k(t)$ in the regime $[2]$.

Theorem 3. [O., 2016]

In the regime $[3]$, we have, as $n \to \infty$,

$$\frac{\beta_{k,n}(t) - \mathbb{E}\{\beta_{k,n}(t)\}}{n^{d/(2\alpha)}} \Rightarrow Z_k(t).$$

Future works

- Combine the following three fields.
  - Heavy Tail Probability
  - Long Range Dependence
  - Topological Data Analysis

Future works

- Combine the following three fields.
  - Heavy Tail Probability
  - Long Range Dependence
  - Topological Data Analysis

Heavy tailed moving average process

- Let (X_i) be an iid $\mathbb{R}^d$-valued random variables with a regularly varying tail.
- Then,

$$Y_n := X_n + AX_{n-1} + BX_{n-2}, \quad n = 1, 2, \ldots$$

defines a MA(2) process, where A and B are non-random $d \times d$-matrices with $\|A\|, \|B\| < 1$.