MATH412/COMPSCI434/MATH713
Fall 2025
Topological Data Analysis

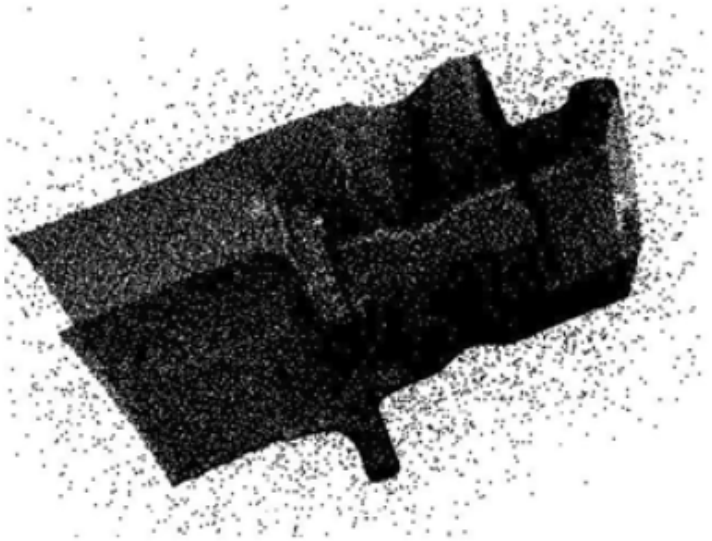
Topic 7: Homology inference, handling of noise, data sparsification

Instructor: Ling Zhou

Handling of Noise

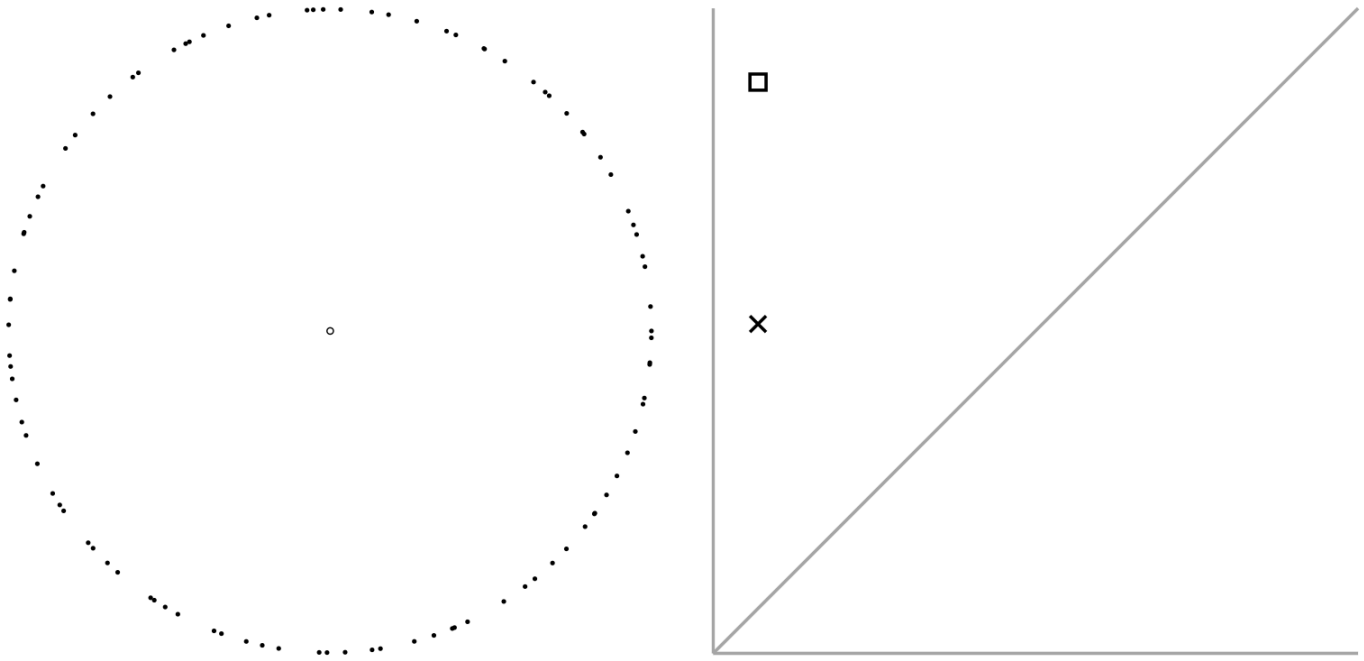
Overview

- Input:
  - A set of points P sampled from a probabilistic measure μ on R^d potentially concentrated on a hidden compact (e.g, manifold) X.
- Goal:
  - Approximate topological features of X
- Previous approach handle Hausdorff type noise
  - Where noise is within a tubular neighborhood of X
  - How about more general noise? e.g. Gaussian noise, background noise, outliers



Courtesy of Chazal et al 2011

Noise

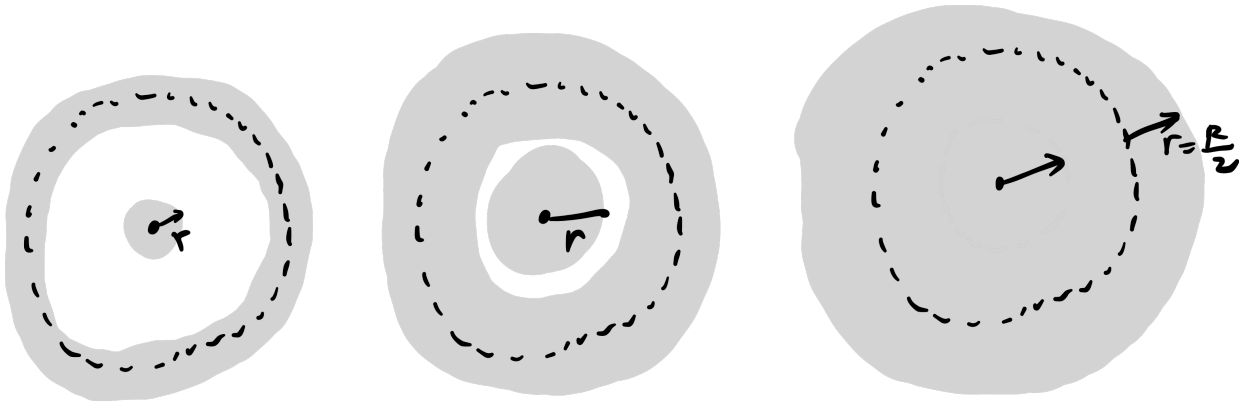


(a) Samples (b) Diagrams

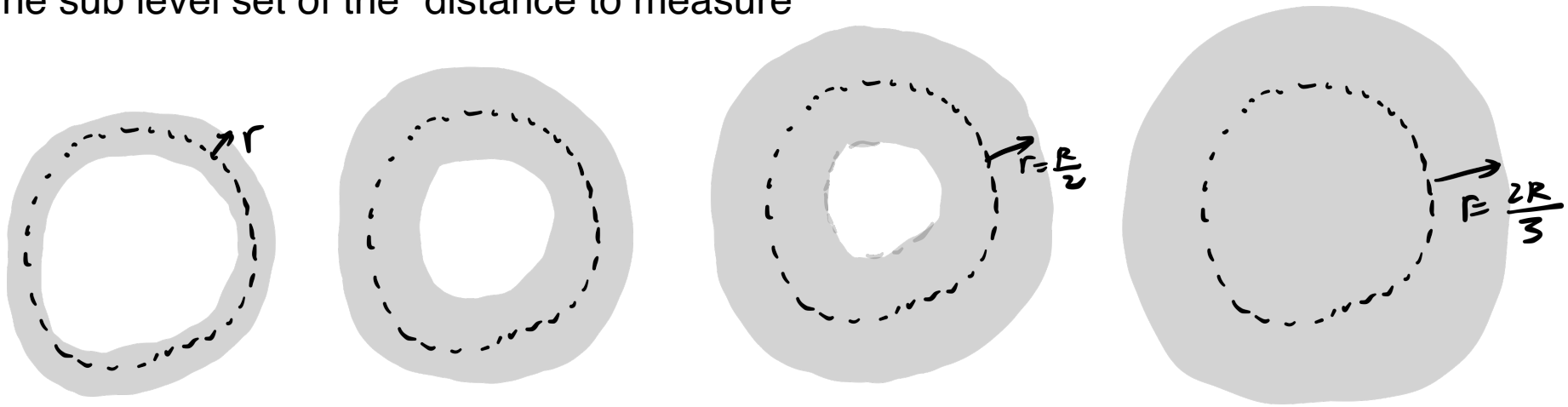
Courtesy of Bendich et al.

Main idea

The sub level set of the distance to set d_X



The sub level set of the “distance to measure”



Key idea:
Use measures to reduce contribution of outliners.

Definitions

A **probability measure** on Ω is a function P from subsets of Ω to the real numbers that satisfies the following axioms:

1. $P(\Omega) = 1$.
2. If $A \subset \Omega$, then $P(A) \geq 0$.
3. If A_1 and A_2 are disjoint, then

$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$

More generally, if $A_1, A_2, \ldots, A_n, \ldots$ are mutually disjoint, then

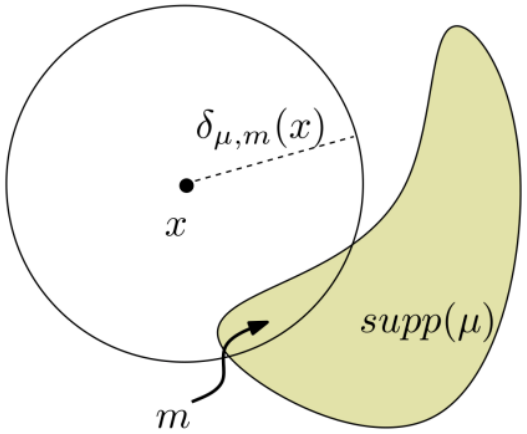
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Example | Given $p \in \mathbb{R}^d$, the Dirac measure at p is

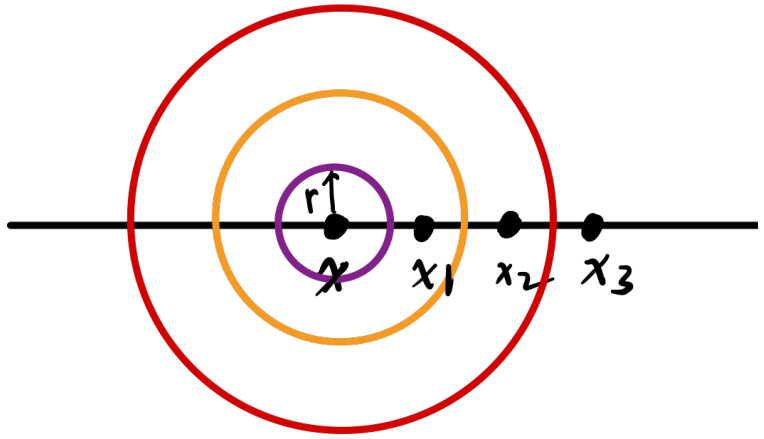
$$\delta_p(A) = \begin{cases} 1, & \text{if } A \ni p \\ 0, & \text{o/w.} \end{cases}$$

Definitions

- μ: a probability measure on $\mathbb{R}^d$; $\mu(\mathbb{R}^d) = 1$
  - Example: Given a set of points $P \subset \mathbb{R}^d$, consider the empirical measure $\mu_P = \frac{1}{n}\sum_{p \in P} \delta_p$ where intuitively every point has measure $\frac{1}{n}$



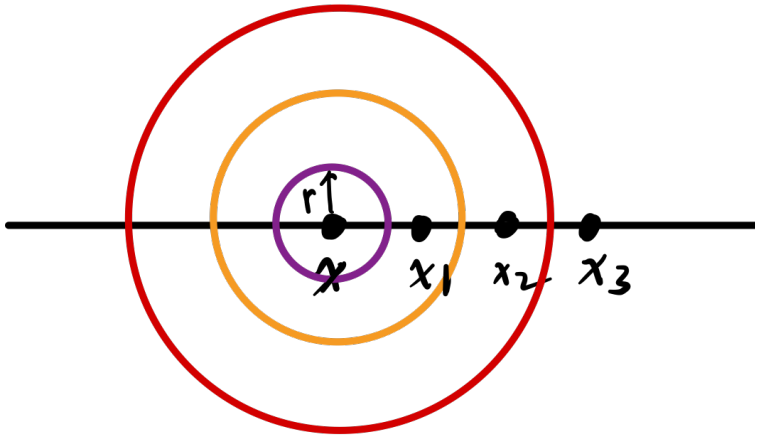
- $\delta_{\mu,m}(\mathrm{x}) := \inf\{\mathrm{r} > 0; \mu\left(\bar{B}(x,\ r)\right) > m\ \}$
  - $\delta_{\mu,m}(\mathrm{x})$ is the radius of the ball necessary in order to enclose mass m



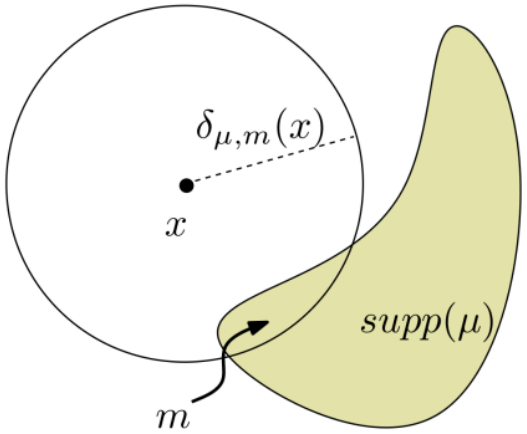
$$\mu = \frac{1}{3}\left(\delta_{x_1} + \delta_{x_2} + \delta_{x_3}\right)$$

$$\delta_{\mu,m}(x) = \begin{cases} \|x - x_1\|, & \text{if } 0 < m \leq \frac{1}{3} \\ \|x - x_2\|, & \text{if } \frac{1}{3} < m \leq \frac{2}{3} \\ \|x - x_3\|, & \text{if } \frac{2}{3} < m \leq 1 \end{cases}$$

Definitions



$$\mu = \frac{1}{3}(\delta_{x_1} + \delta_{x_2} + \delta_{x_3})$$



Distance to Measures (DTM)

- *Distance to measure* d_{μ,m_0}:

$$d^2_{\mu,m_0}(x) = \frac{1}{m_0}\int_0^{m_0} \delta_{\mu,m}(x)^2\, dm$$

  - d_{μ,m_0}(x) averages distance within a range and is more robust to noise.
  - d_{μ,m_0} depends on a mass parameter m_0

- Given a set of points $P \subset \mathbb{R}^d$, consider $\mu_P = \frac{1}{n}\sum_{p\in P}\delta_p$. For $m_0 = k/n$,

$$d_{\mu_P,m_0}(x) = \sqrt{\frac{1}{k}\sum_{q\in kNN(x)} \|x-q\|^2} =: d_{\mu_P,k}(x),$$

Note: $d_{\mu_P,1} = d_X$

  - where $kNN(x)$ denotes the set of k nearest neighbors of x in P

(Read Def. 1.18 of textbook.) isotopy ⇒ homeomorphism.

Properties

- Theorem [Distance-Likeness] [Chazal et al 2011]

  d_{μ,m_0} is distance like. That is:

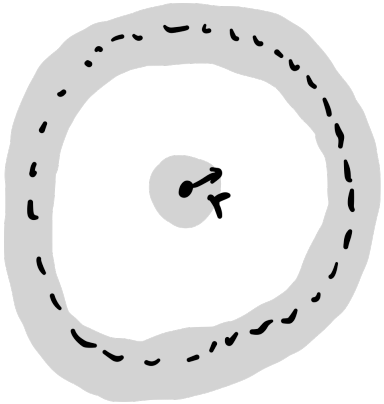
  - The function d_{μ,m_0} is 1-Lipschitz.
  - The function d^2_{μ,m_0} is 1-semiconcave, meaning that the map $x \rightarrow d^2_{\mu,m_0}(x) - \|x\|^2$ is concave.

- Theorem [Isotopy Lemma]:

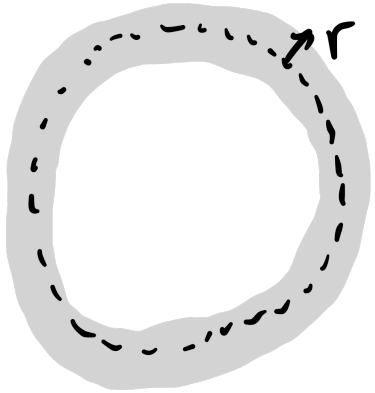
  Let ϕ be a distance-like function and $r_1 < r_2$ be two positive numbers such that ϕ has no critical points in the subset $\phi^{-1}\Big(\big[r_1, r_2\big]\Big)$. Then all the sublevel sets $\phi^{-1}\Big(\big[0, r\big]\Big)$ are isotopic for $r \in [r_1, r_2]$.

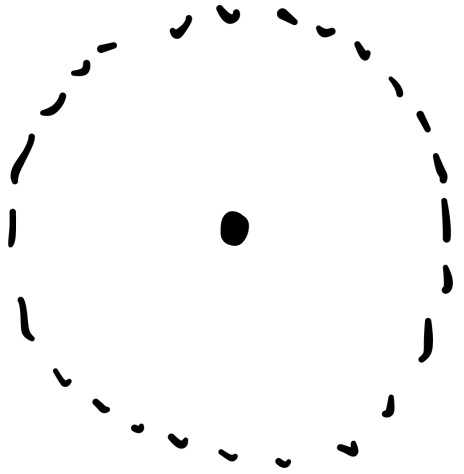
Sublevel set of distance to measure

$$\sqrt{\frac{1}{k}\sum_{q\in kNN(x)}\|x-q\|^2} =: d_{\mu_P,k}(x)$$



$d_X^{-1}((-\infty, r])$





$d_{\mu,k}^{-1}((-\infty, r])$ depending on k.

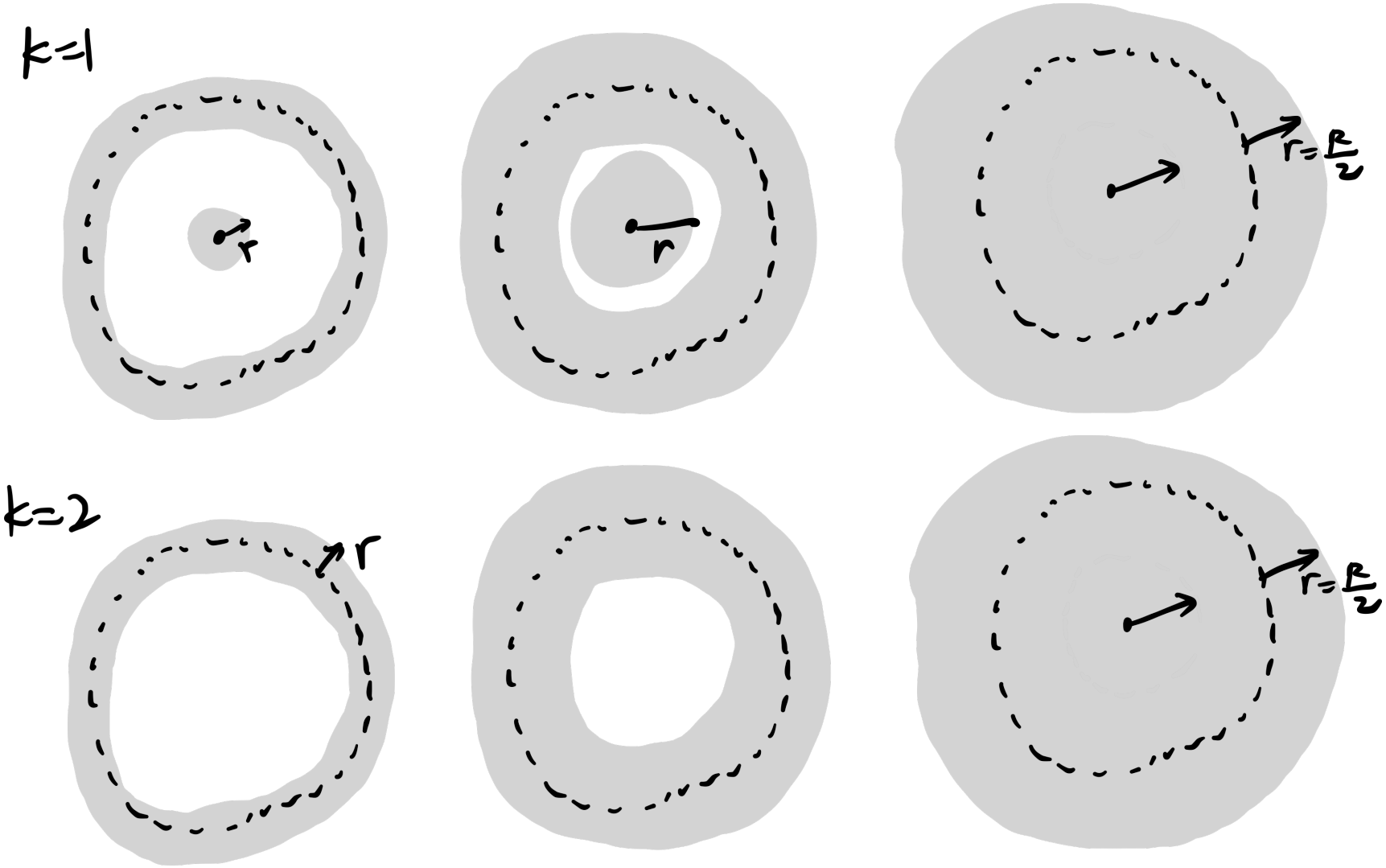
(1) $d_{\mu,1}: x \longmapsto$ distance between x & its closest pt.

$d_{\mu,1}^{-1}((-\infty, r]) = \{x \mid$ distance between x & its closest pt $\leq r\}$

(2) $d_{\mu,k}^{-1}((-\infty, r]) = \{x \mid$ ave dist between x & its k NN pt $\leq r\}$

Sublevel set of distance to measure

$$\sqrt{\frac{1}{k}\sum_{q\in kNN(x)}\|x-q\|^2} =: d_{\mu_P,k}(x)$$



Stability of DTM

- **Theorem [Stability] [Chazal et al 2011]**

  Let $\mu,\ \mu'$ be two probability measures on $\mathbb{R}^d$ and $m_0 > 0$. Then

  $$\left\| d_{\mu,\, m_0} - d_{\mu', m_0} \right\|_\infty \leq \frac{1}{\sqrt{m_0}} W_2(\mu,\ \mu')$$

- **Theorem [Stability of PD] [Buchet et al 2016]**

  Let $\mu,\ \mu'$ be two probability measures on $\mathbb{R}^d$ and $m_0 > 0$. Then,

  $$d_B(Dgm(d_{\mu,m_0}), Dgm(d_{\mu',m_0})) \leq \|d_{\mu,m_0} - d_{\mu,m_0}\|_\infty \leq \frac{1}{\sqrt{m_0}} W_2(\mu, \mu')$$

Relation to Distance Function

- Let ν_X denote the uniform measure on a manifold X

  - Theorem [Approximation Distance]:

    $\|d_X - d_{\nu_X, m_0}\|_\infty \leq C(X)^{-1/d} m_0^{1/d}$

    where X is a d-dimensional smooth manifold and $C(X)$ is a quantity depending on X and d

- In particular, $\lim_{m \to 0} d_{\nu_X, m} = d_X$

Topological inference

Corollary 4.11 *Let μ be a measure and K its support. Suppose that μ has dimension at most k and that $\mathrm{reach}_\alpha(d_K) \geq R$ for some $R > 0$. Let μ' be another measure, and let ε be an upper bound on the uniform distance between d_K and d_{μ',m_0}. Then, for any $r \in [4\varepsilon/\alpha^2, R - 3\varepsilon]$, the r-sublevel sets of d_{μ',m_0} and the offsets K^η, for $0 < \eta < R$, are homotopy equivalent, as soon as*

$$\mathrm{W}_2(\mu, \mu') \leq \frac{R\sqrt{m_0}}{5 + 4/\alpha^2} - C(\mu)^{-1/k} m_0^{1/k+1/2}.$$

Noisy sample

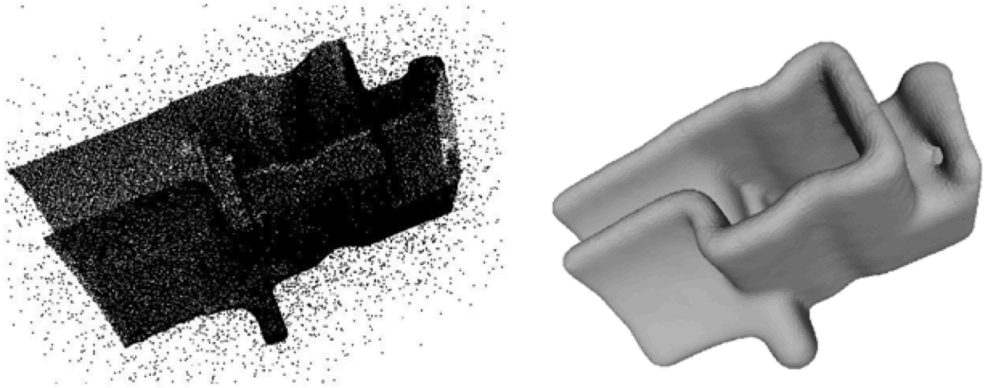
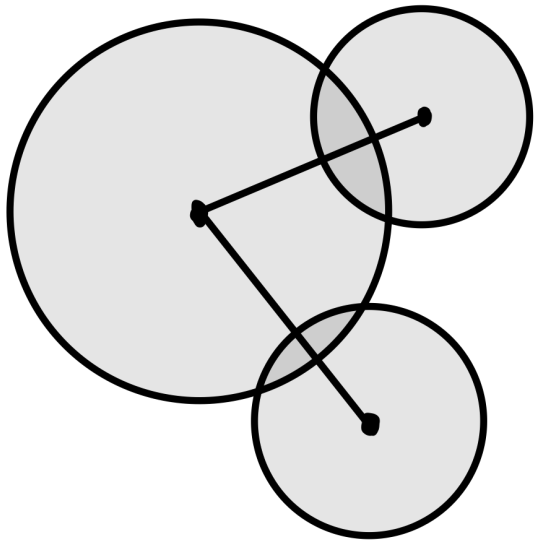


Fig. 1 *Left*: a point cloud sampled on a mechanical part to which 10% of outliers (uniformly sampled in a box enclosing the model) have been added. *Right*: the reconstruction of an isosurface of the distance function d_{μ_C,m_0} to the uniform probability measure on this point cloud

How to compute/approximate the distance to measures?

- Previous theorem suggests that using distance-to-measure, instead of distance can be used to approximate topology.
  - But the sublevel set of $d_{\mu,m}$ (the DTM filtration) is hard to compute directly.
  - In practice, use union of weighted balls of non-uniform radius to approximate
- One way to achieve this is through the power-distance [Buchet et al., 2016]



Power distance

Definition 4.1. *Given a metric space* $\mathbb{X}$*, a set* P *and a function* $w : P \to \mathbb{R}$*, we define the* power distance f *associated with* (P, w) *as*

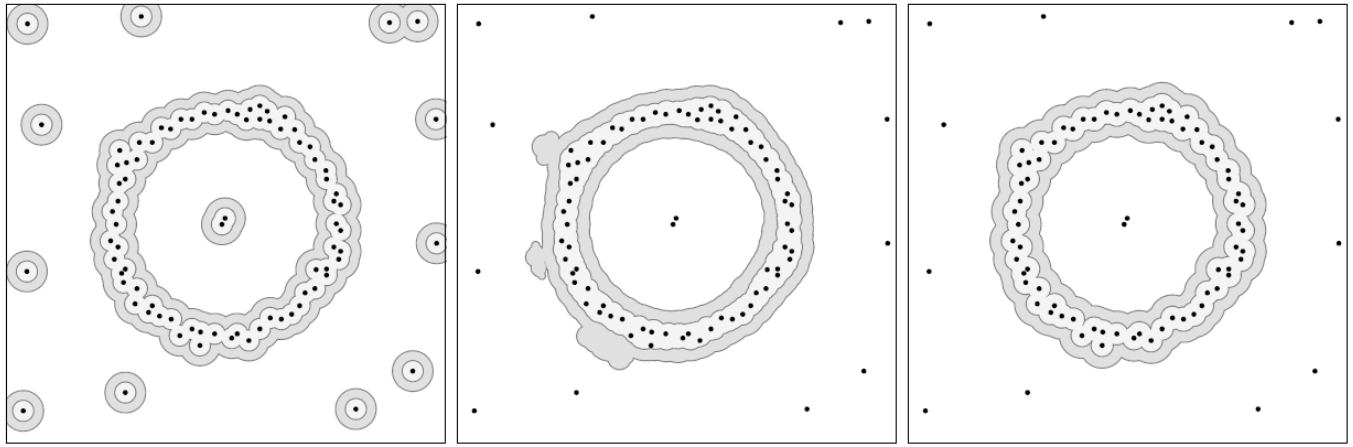
$$f(x) = \sqrt{\min_{p \in P} d_{\mathbb{X}}(p, x)^2 + w_p^2},$$

where w_p *is the value of* w *at the point* p*.*

- Note when $w_p \equiv 0$, then $f(x) = d_P(x)$
- Let $r_p(\alpha) = \sqrt{\alpha^2 - w_p^2}$. Then, $f^{-1}((-\infty, \alpha]) = \cup_{p \in P} B(p, r_p(\alpha))$
- Set $w_p = d_{\mu,m}(p)$.
  - The corresponding f is an approximation of $d_{\mu,m}$
  - The sub level set filtration of f is an approximation of the one for $d_{\mu,m}$

Weighted Rips Filtration

- Weighted Rips filtration
  - $wR^{\alpha}(P) = \{\sigma = \left(p_{i_0}, \ldots, p_{i_s}\right) \mid d\left(p_{i_j}, p_{i_{j'}}\right) \leq r_{p_{i_j}}(\alpha) + r_{p_{i_{j'}}}(\alpha), \ \forall j \neq j' \in [0, s]\}$
  - $w\mathcal{F}: \ wR^{\alpha_0}(P) \hookrightarrow wR^{\alpha_1}(P) \hookrightarrow \cdots \hookrightarrow wR^{\alpha_m}(P)$
- The persistence diagram induced by $w\mathcal{F}$, which is an approximation of $Dgm(d_{\mu,m})$



DTM
||

weighted Rips ⇝ weighted Čech ⇝ sublevel of f ⇝ sublevel of $d_{\mu,m}$

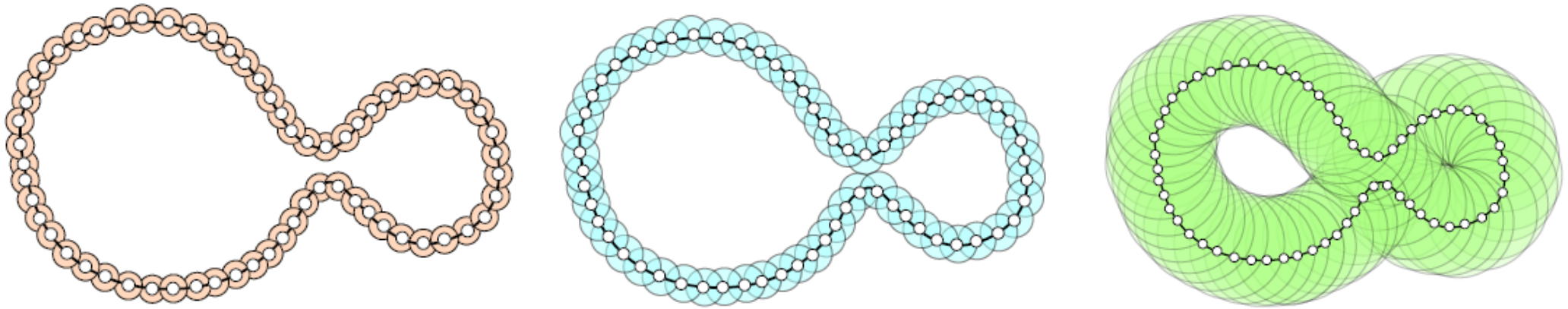
Additional reading

- Youtube videos

Data Sparsification

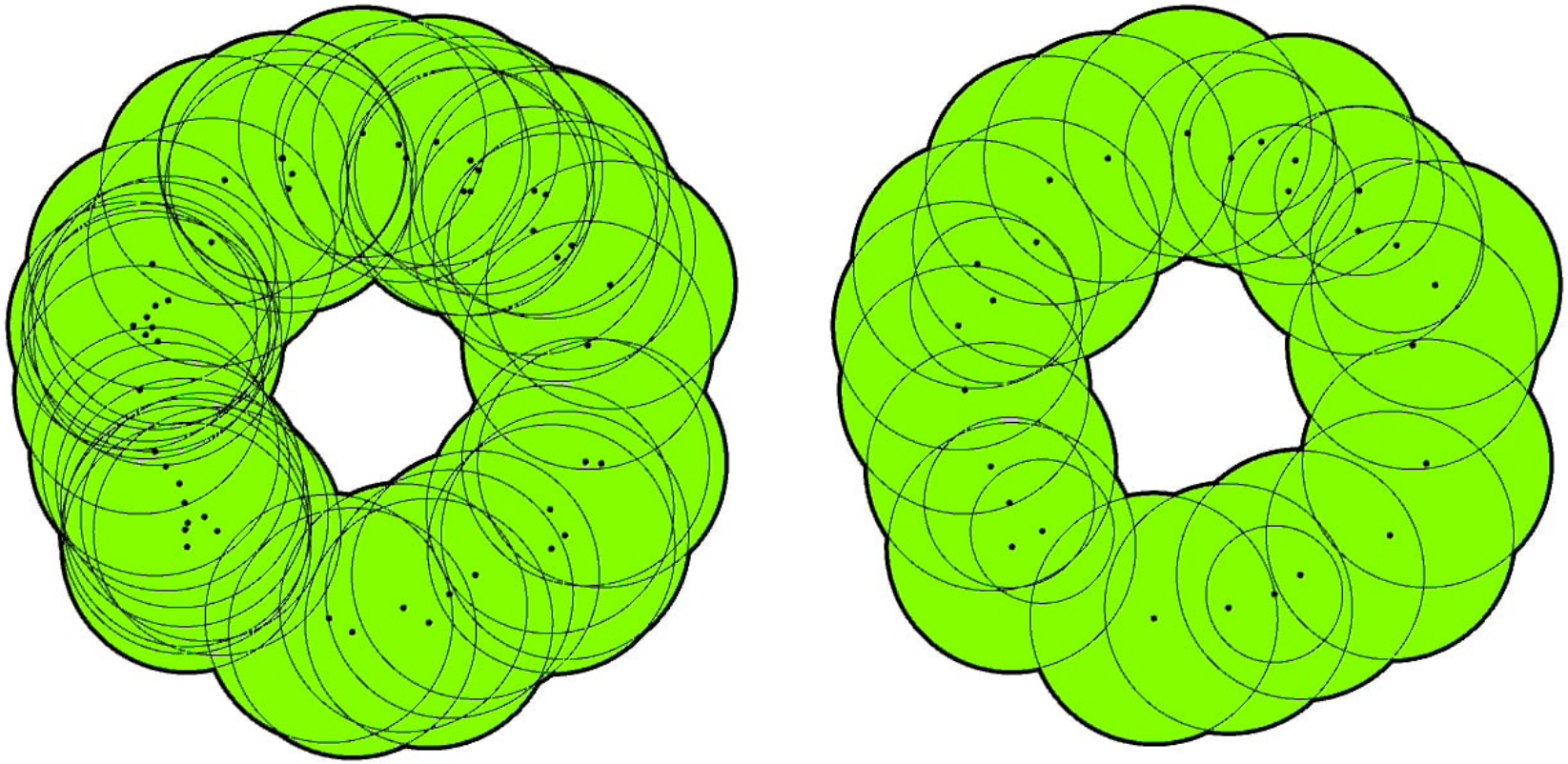
Rips Filtration

- Size becomes huge quickly
- But not all simplices are needed, especially at large scales!
- Idea:
  - Use fewer number of points (balls) at larger scales



Additional reading

- A video explanation by *Cavanna, Jahanseir and Sheehy*



A good subsample

- Given a set of points P, a subset $Q \subseteq P$ is a ε-net of P if
  - (covering-condition): Q is a ε-**dense**, i.e., for any $p \in P$, $d(p, Q) \leq \epsilon$, i.e., $d_H(P, Q) \leq \epsilon$
  - (sparsity-condition): Q is a ε-**sparse**, i.e., for every $q \neq q' \in Q$, $d\left(q,\ q'\right) \geq \varepsilon$

- Covering-condition guarantees that Q represents P well at scale ε
- Sparsity-condition makes sure Q is not overly dense, thus is of small cardinality

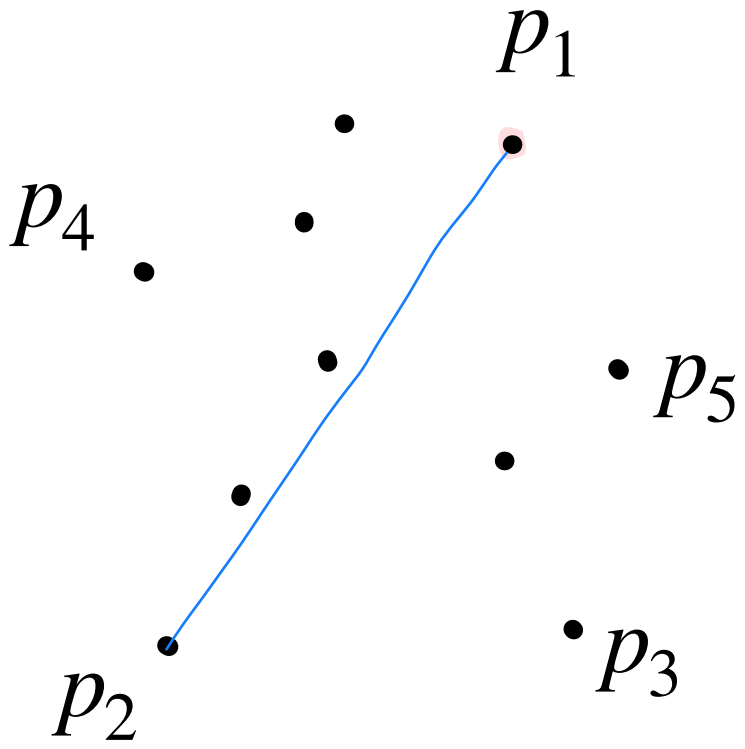
How to generate an ϵ-net?

- Given input point set $P \subset R^d$
- Let $\{p_1, p_2, \ldots, p_n\}$ be the sequence of points obtained via *farthest point sampling procedure*
  - Pick arbitrary point $p_1 \in P$ and set $P_1 = \{p_1\}$
  - Pick p_i recursively as $p_i = argmax_{p \in P \setminus P_{i-1}} d(p, P_{i-1})$. Set $P_i = P_{i-1} \cup \{p_i\}$
  - Easy to see that $P = P_n \supset P_{n-1} \supset \cdots \supset P_2 \supset P_1$
- Exit-time of $p = p_i$ is set to be $t_{p_i} := d(p_i, P_{i-1})$

Net-tower

$p_i = argmax_{p \in P \setminus P_{i-1}} d(p, P_{i-1})$.

- Exit-time of $p = p_i$ is set to be $t_{p_i} := d(p_i, P_{i-1})$
- Each P_i is a t_{p_i}- net of P



(1) P_1: arbitrary & $P_1 = \{p_1\}$

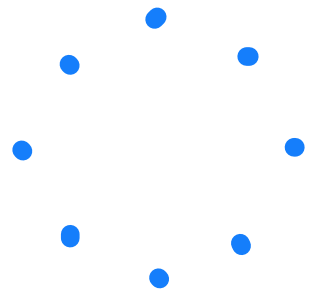
(2) P_2: farthest pt to P_1

- $P_2 = P_1 \cup \{p_2\}$
- $t_{p_2} = d(p_2, P_1)$

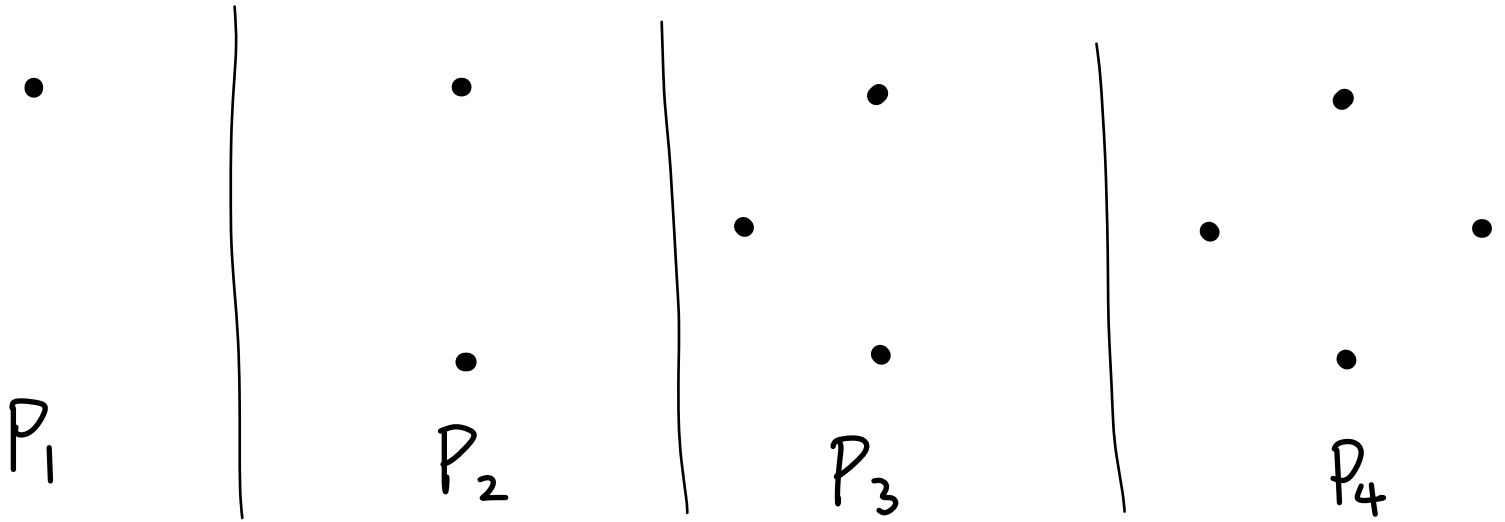
(3) P_3: farthest pt to P_2

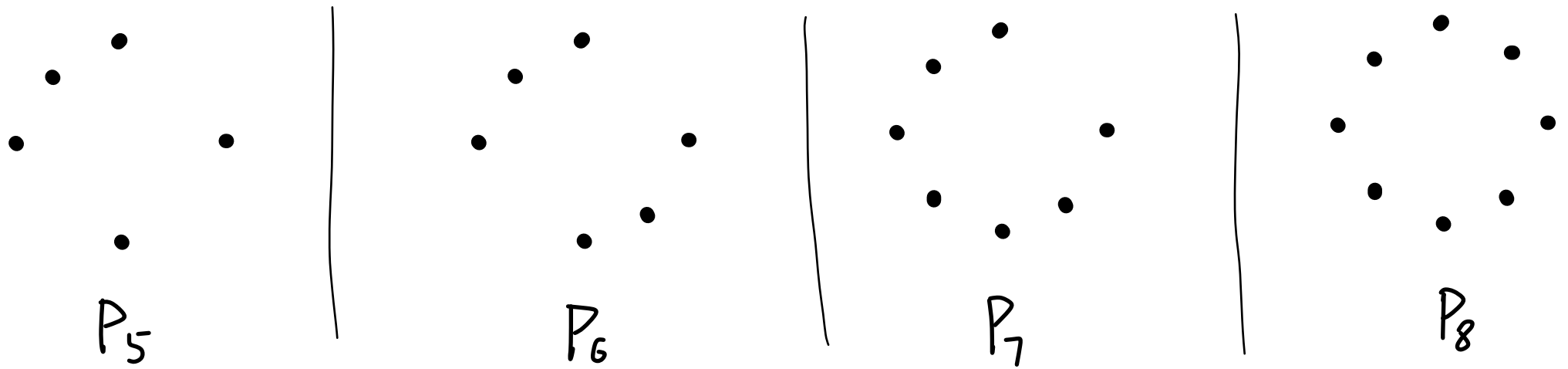
- $P_3 = P_2 \cup \{p_3\}$
- $t_{p_3} = d(p_3, P_2)$

Example



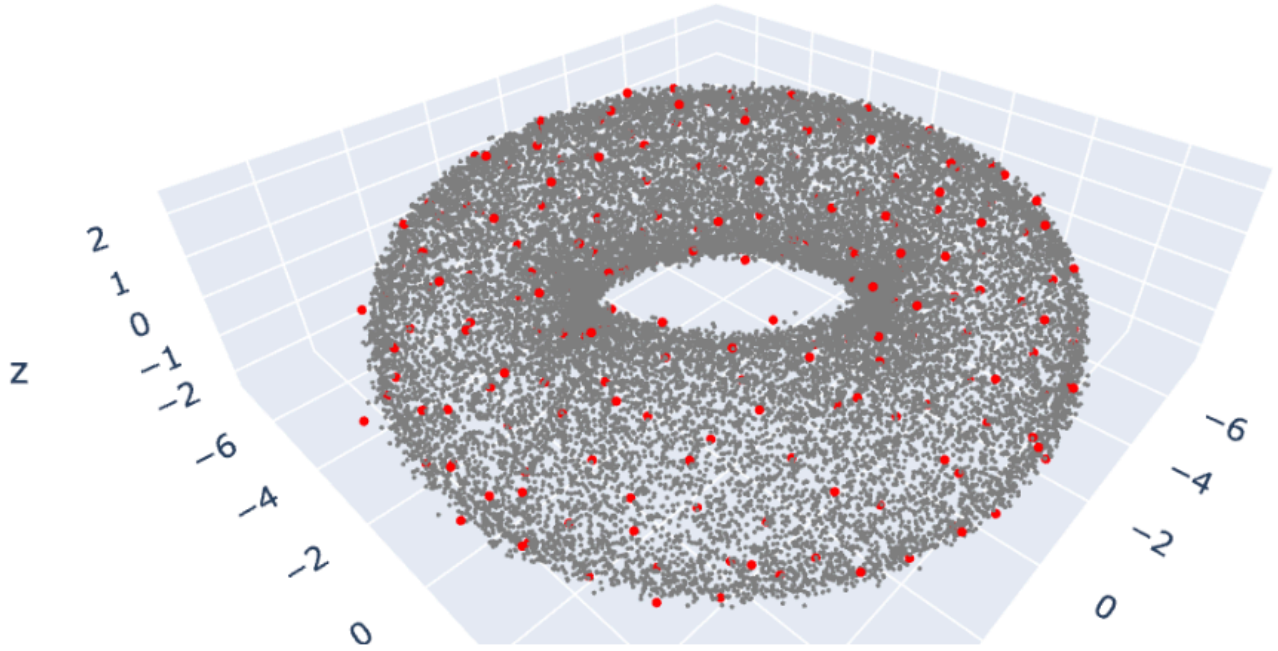
8 equally distributed pts on circle





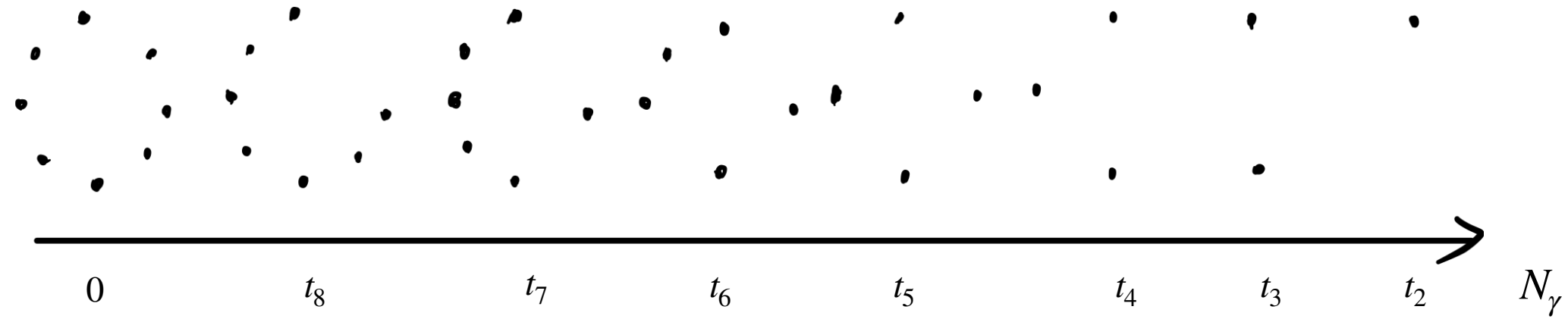
Filtration Choice 1

- Choose an ϵ-net Q of P, and use PH of Q to approximate PH of P.

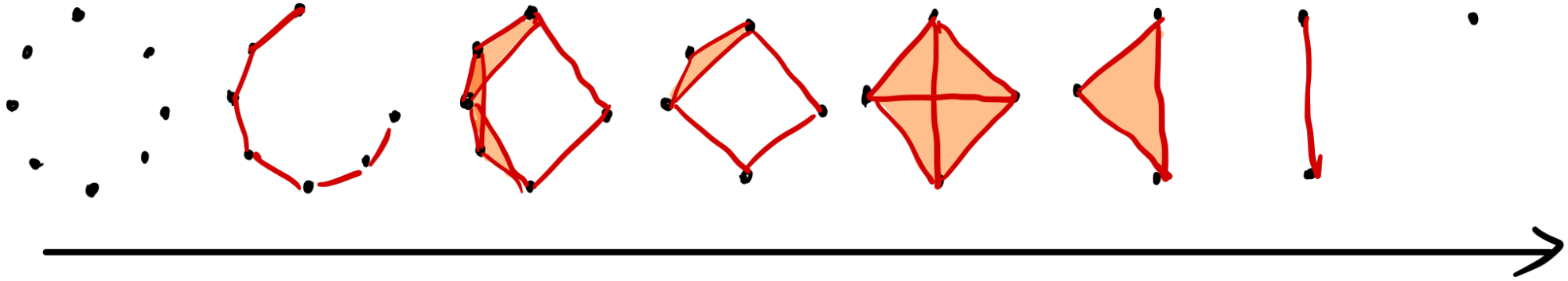


Filtration Choice 2

- Index by exit time $t_i := d(p_i, p_{i-1})$

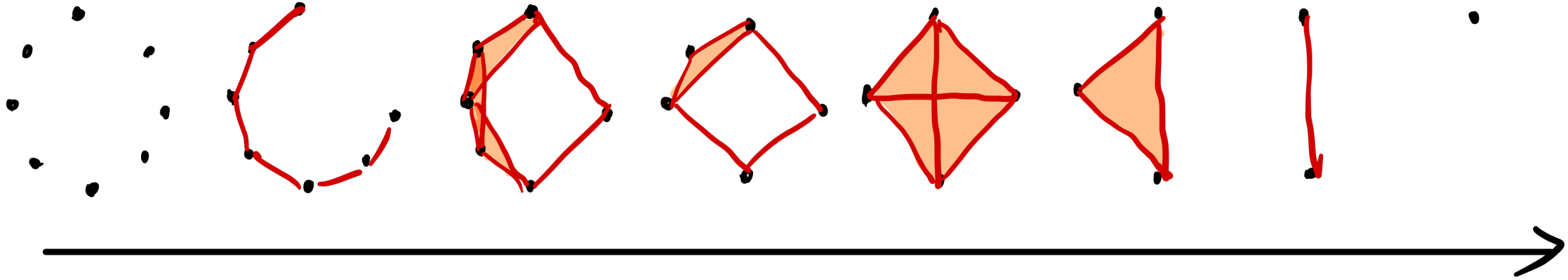


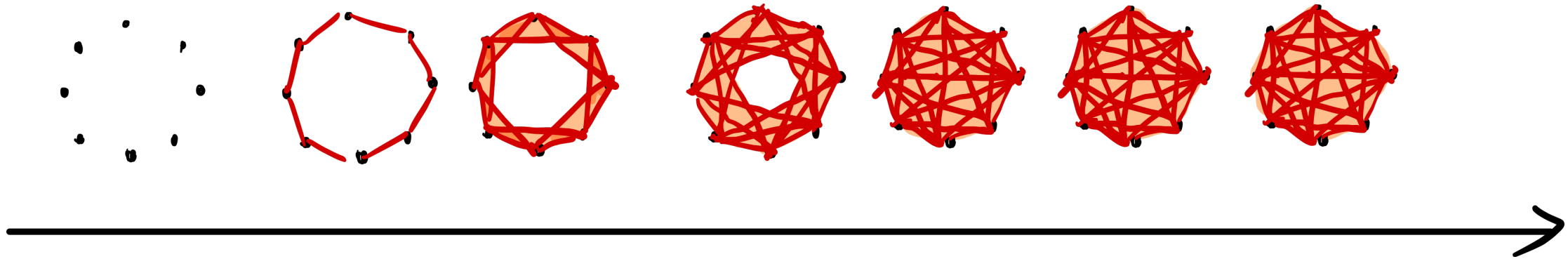
- Apply Rips construction $VR_{t_i}(p_{i-1})$



Naive Rips construction does not work well

A Sparse Rips filtration



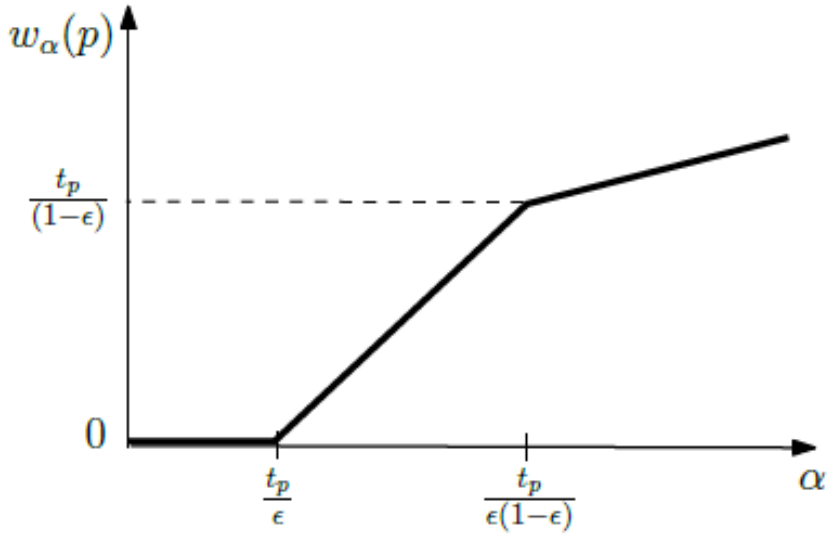


Original Rips filtration

Weights and weighted distance

- Using exit-time, we assign a weight $w_p(\alpha)$ for each point p at scale α

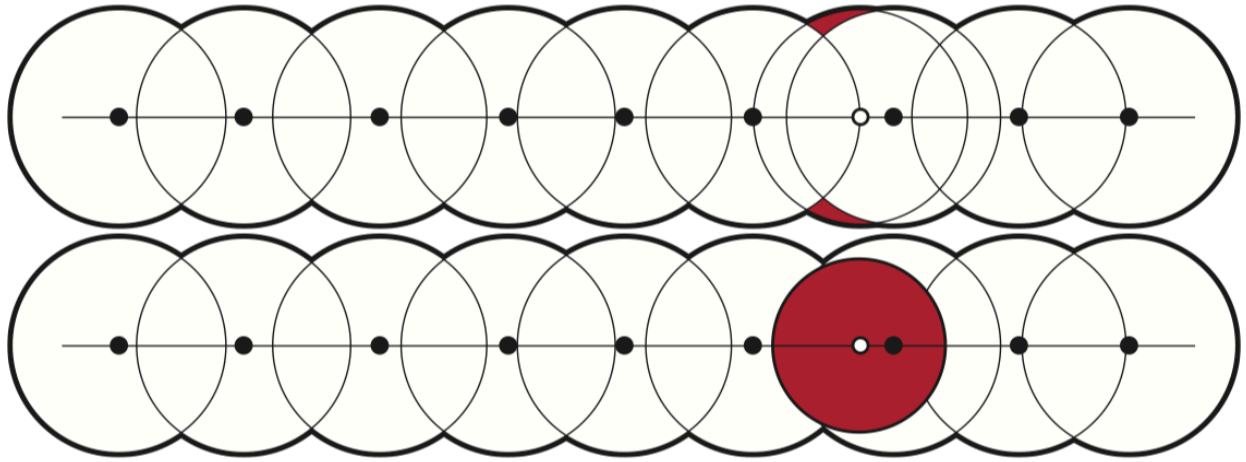
$$w_p(\alpha) = \begin{cases} 0 & \frac{t_p}{\varepsilon} \geq \alpha \\ \alpha - \frac{t_p}{\varepsilon} & \frac{t_p}{\varepsilon} < \alpha < \frac{t_p}{\varepsilon(1-\varepsilon)} \\ \varepsilon\alpha & \frac{t_p}{\varepsilon(1-\varepsilon)} \leq \alpha \end{cases}$$



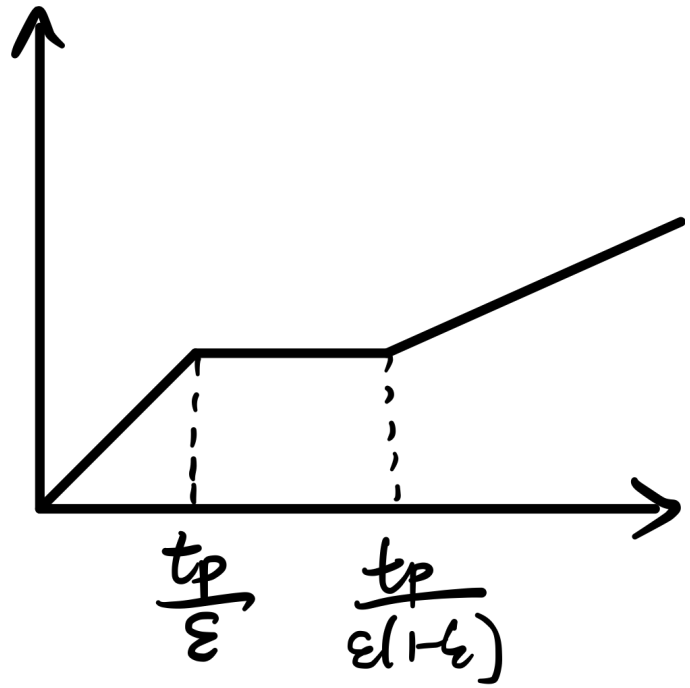
- Net-induced distance at scale α is:
  - $\hat{d}_\alpha(p,\ q) = d(p,\ q) + w_p(\alpha) + w_q(\alpha)$

Weights and weighted distance

- Let $r_p(\alpha) = \alpha - w_p(\alpha)$
- $\hat{d}_\alpha(p,q) = d(p,q) + w_p(\alpha) + w_q(\alpha) \leq 2\alpha$ means two balls $B(p, r_p(\alpha))$ and $B(q, r_q(\alpha))$ have intersection

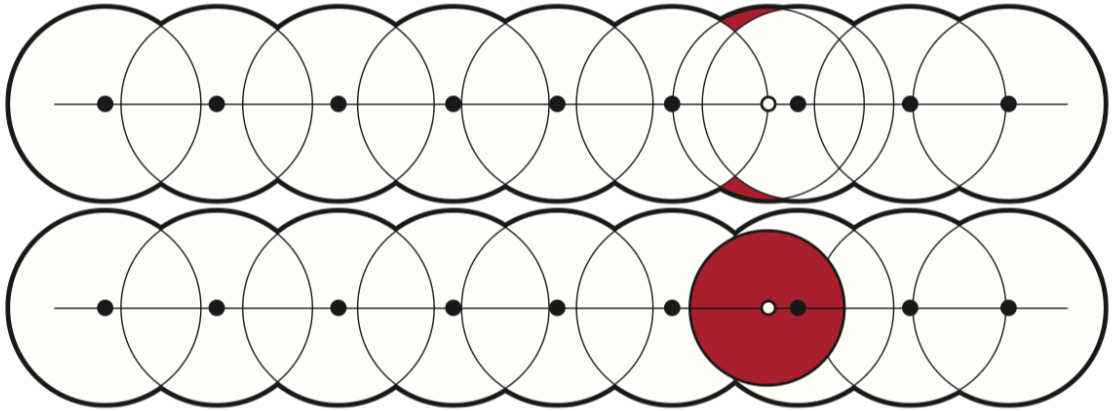


Courtesy of Sheehy 2012



Sparse Rips complexes

- Let $\hat{VR}^{\alpha}(P) = \{\sigma \subset P : \hat{d}_{\alpha}(p,q) \leq 2\alpha \text{ for } p,q \in \sigma\}$
- $VR^{\alpha(1-\epsilon)}(P) \subseteq \hat{VR}^{\alpha}(P) \subseteq VR^{\alpha}(P)$
  - $\hat{d}_{\alpha}(p,q) = d(p,q) + w_p(\alpha) + w_q(\alpha) \leq 2\alpha(1-\epsilon) + 2\epsilon\alpha = 2\alpha$
  - $d(p,q) \leq \hat{d}_{\alpha}(p,q) \leq 2\alpha$
- But $\hat{VR}^{\alpha}(P)$ is not necessarily nested: it may not form a filtration



Sparse Rips complexes

$N_\gamma, \overline{N}_\gamma \subset P$

Definition 6.5 (Sparse (Vietoris-)Rips)**.** Given a set of points $P \subset \mathbb{R}^d$, a constant $0 < \varepsilon < 1$, and the open net-tower $\{N_\gamma\}$ as well as the closed net-tower $\{\overline{N}_\gamma\}$ for P as introduced above, the *open sparse-Rips complex at scale* α is defined as

$$\mathsf{Q}^\alpha := \{\sigma \subseteq N_{\varepsilon(1-\varepsilon)\alpha} \mid \forall p, q \in \sigma,\ \widehat{\mathsf{d}}_\alpha(p,q) \leq 2\alpha\} = \hat{V}R^\alpha\left(N_{\varepsilon(1-\varepsilon)\alpha}\right) \quad (6.9)$$

while the *closed sparse-Rips at scale* α is defined as

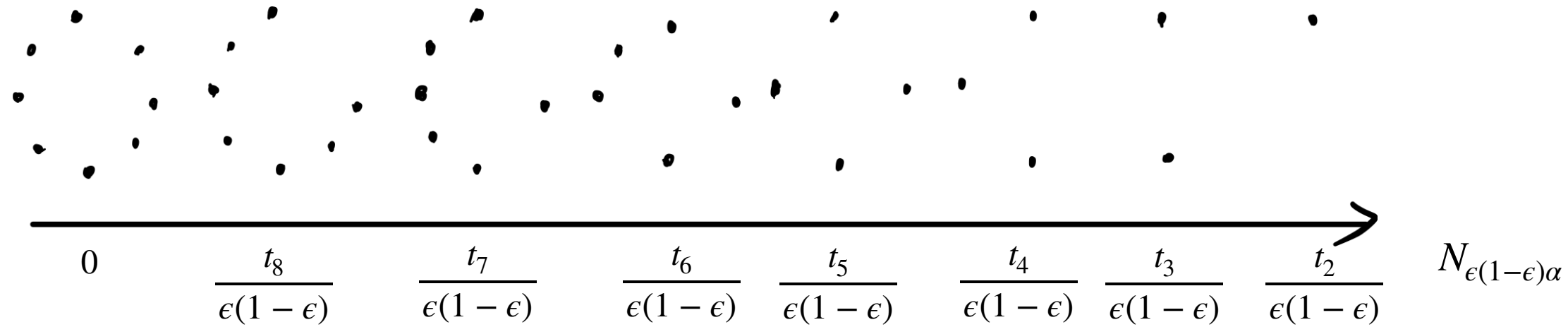
$$\overline{\mathsf{Q}}^\alpha := \{\sigma \subseteq \overline{N}_{\varepsilon(1-\varepsilon)\alpha} \mid \forall p, q \in \sigma,\ \widehat{\mathsf{d}}_\alpha(p,q) \leq 2\alpha\}. = \hat{V}R^\alpha\left(\overline{N}_{\varepsilon(1-\varepsilon)\alpha}\right) \quad (6.10)$$

Set $\mathsf{S}^\alpha := \cup_{\beta \leq \alpha} \overline{\mathsf{Q}}^\alpha$, which we call the *cumulative complex at scale* α. The *(ε-)sparse Rips filtration* then refers to the $\mathbb{R}$-indexed filtration $\mathcal{S} = \{\mathsf{S}^\alpha \hookrightarrow \mathsf{S}^\beta\}_{\alpha \leq \beta}$.

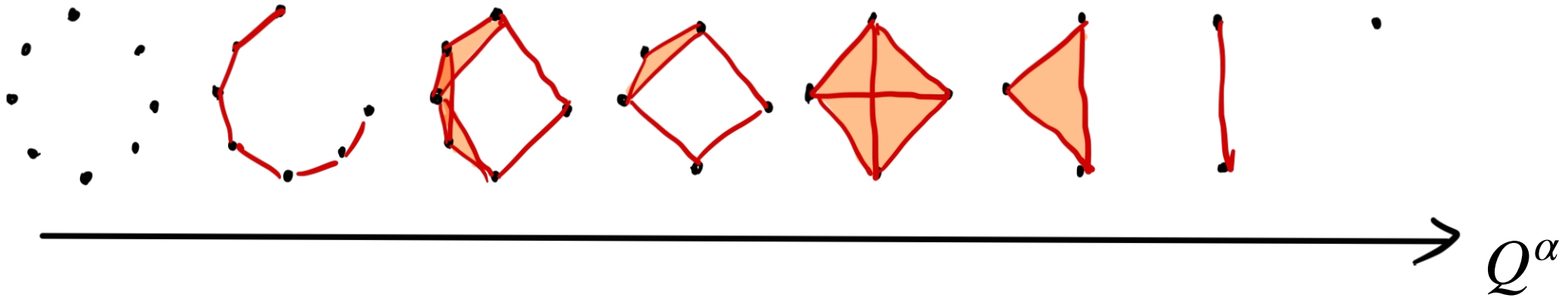
- Larger ϵ corresponds to sparser complexes

Filtration Choice 3

- Index by scaled exit time $\frac{t_i}{\epsilon(1-\epsilon)}$



- Apply weighted Rips construction $\hat{VR}^{\frac{t_i}{\epsilon(1-\epsilon)}}(P_{i-1})$



Guarantee of Sparse Rips Filtration

Theorem 6.4. *Let $P \subset \mathbb{R}^d$ be a set of n points where d is a constant, and $\mathcal{R}(P) = \{\mathbb{VR}^r(P)\}$ be the Vietoris-Rips filtration over P. Given net-towers $\{N_\gamma\}$ and $\{\overline{N}_\gamma\}$ induced by exit-times $\{\mathrm{t}_p\}_{p\in P}$, let $\mathcal{S}(P) = \{\mathsf{S}^\alpha\}$ be its corresponding ε-sparse Rips filtration as defined in Definition 6.5. Then, for a fixed $0 < \varepsilon < \frac{1}{3}$,*

(i) $\mathcal{S}(P)$ and $\mathcal{R}(P)$ are multiplicatively $\frac{1}{1-\varepsilon}$-interleaved at the homology level. Thus, for any $k \geq 0$, the persistence diagram $\mathrm{Dgm}_k\mathcal{S}(P)$ is a $\log\frac{1}{1-\varepsilon}$-approximation of $\mathrm{Dgm}_k\mathcal{R}(P)$ at the log-scale.

(ii) For any fixed dimension $k \geq 0$, the total number of k-simplices ever appeared in $\mathcal{S}(P)$ is $\Theta((\frac{1}{\varepsilon})^{kd}n)$. v.s. n^k in Rips

$d_B(\mathrm{Dgm}(\text{sparse Rips}), \mathrm{Dgm}(\text{Rips})) \leq \log\frac{1}{1-\varepsilon}$

- Finally, this sparsification strategy can be extended to handle weighted Rips filtration w.r.t. distance to measures.
  - [Buchet, Chazal, Oudot, Sheehy, CGTA 2016]

- An implementation in Ripser
- Giotto-TDA

Quantum computing

- Any simplicial complex on an n point set has at most 2^n simplices
- One can possibly use n qubits and superposition to represent these simplicial complexes
- Ameneyro et al 2024

Quantum persistent homology

Bernardo Ameneyro[1] · Vasileios Maroulas[1] · George Siopsis[2]

Received: 18 April 2022 / Revised: 26 October 2023 / Accepted: 28 December 2023