

DSC214

Topological Data Analysis

Topic 6: Homology inference, denoising, data sparsification

Instructor: Zhengchao Wan

Today

- ▶ Homology inference
- ▶ Handling of noise
- ▶ Data sparsification

Section 1: Topology inference from PCD and manifolds

▶ Input:

- ▶ A set of points $P \subseteq R^d$ sampled on/around X

▶ Question:

- ▶ How to approximate the persistence module induced by F_X ?

Target filtration (F_X): $X^{r_0} \subseteq X^r \subseteq \dots X^r \subseteq \dots$

Intermediate filtration: $P^{r_0} \subseteq P^{r_1} \subseteq \dots P^r \subseteq \dots$

\approx
By Nerve Lemma

Čech filtration (\mathcal{C}_X): $C^{r_0} \subseteq C^{r_1} \subseteq \dots C^r \subseteq \dots$

\approx
Two sequence interleave

Rips filtration (\mathcal{R}_X): $R^{r_0} \subseteq R^{r_1} \subseteq \dots R^r \subseteq \dots$

Hausdorff distance between subsets

- ▶ If $P \subseteq X$ satisfies that $d_H(P, X) = \inf\{r : X \subseteq P^r\} < \epsilon$

Target filtration (F_X): $X^{r_0} \subseteq X^{r_1} \subseteq \dots X^r \subseteq \dots$

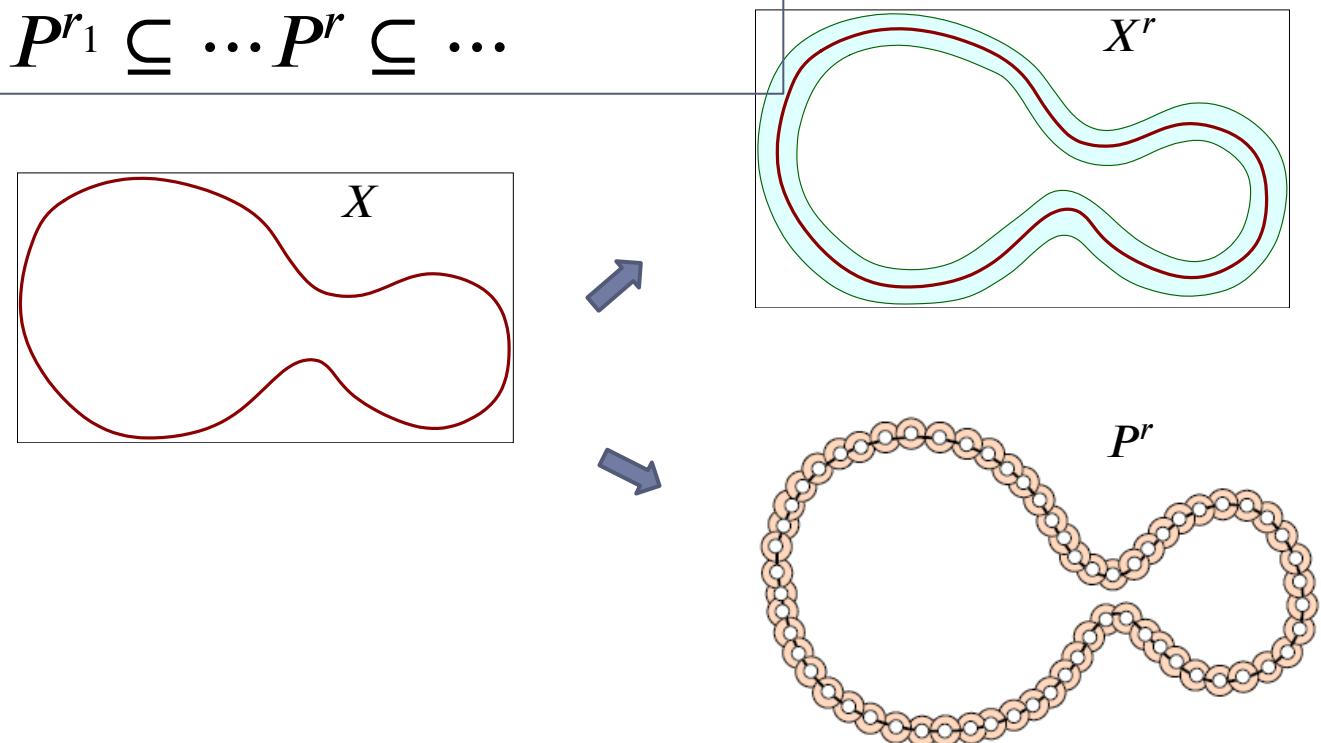
Intermediate filtration: $P^{r_0} \subseteq P^{r_1} \subseteq \dots P^r \subseteq \dots$

- ▶ Note that

- ▶ $P^r \subset X^{r+\epsilon}$

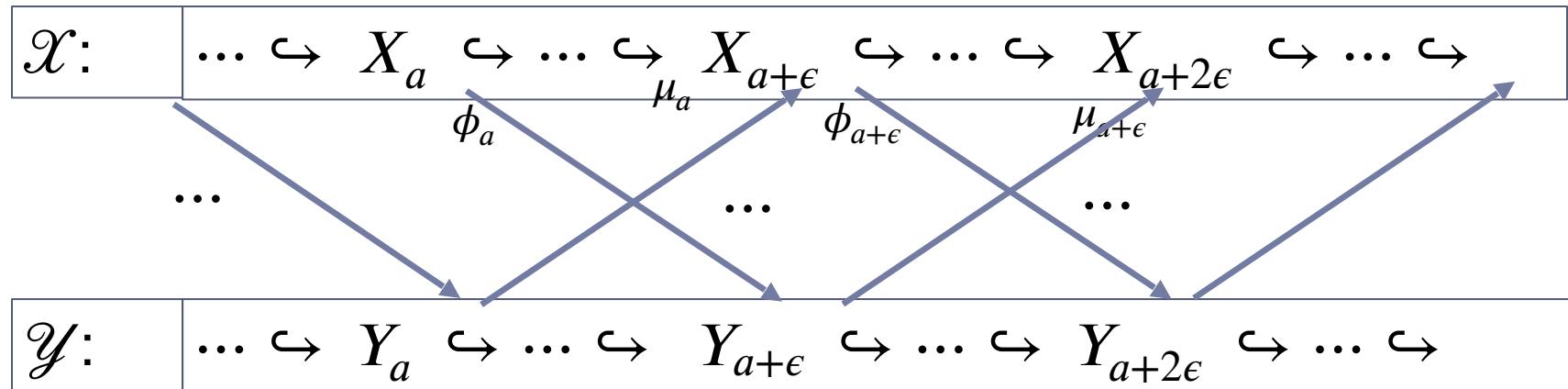
- ▶ $X^r \subset P^{r+\epsilon}$

- ▶ So $d_I(P, F_X) \leq \epsilon$



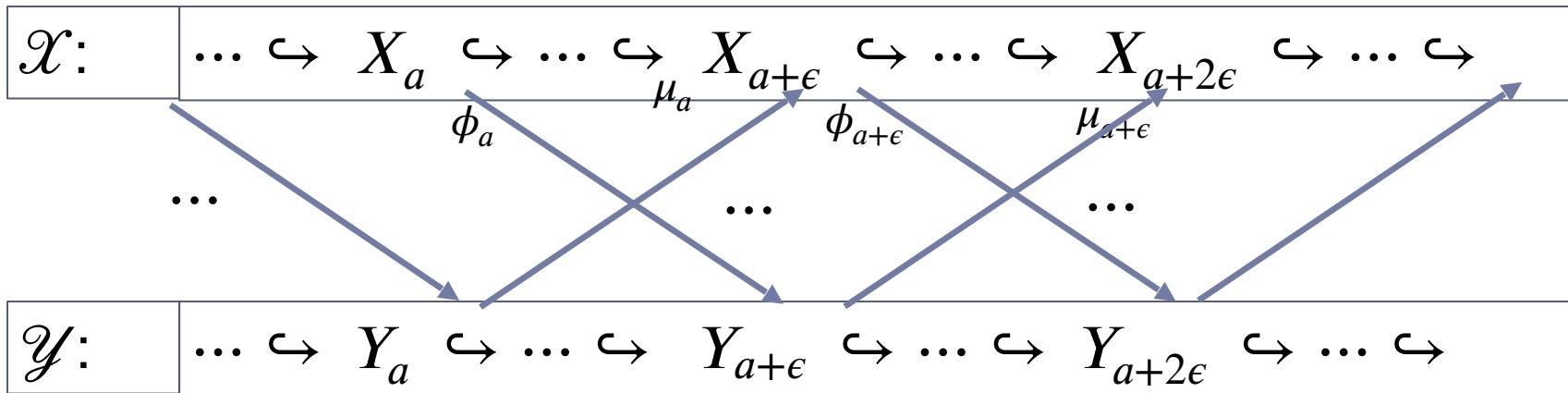
Recall: Interleaving filtrations

- Given two filtrations \mathcal{X} and \mathcal{Y} ϵ -interleaves if the following diagram commutes.



Recall: Interleaving filtrations

- Given two filtrations \mathcal{X} and \mathcal{Y} ϵ -interleaves if the following diagram commutes.



Theorem

Given two ϵ -interleaved filtrations \mathcal{X} and \mathcal{Y} , let D_X and D_Y be the corresponding persistence diagrams induced by them. We then have:

$$d_B(D_X, D_Y) \leq \epsilon$$

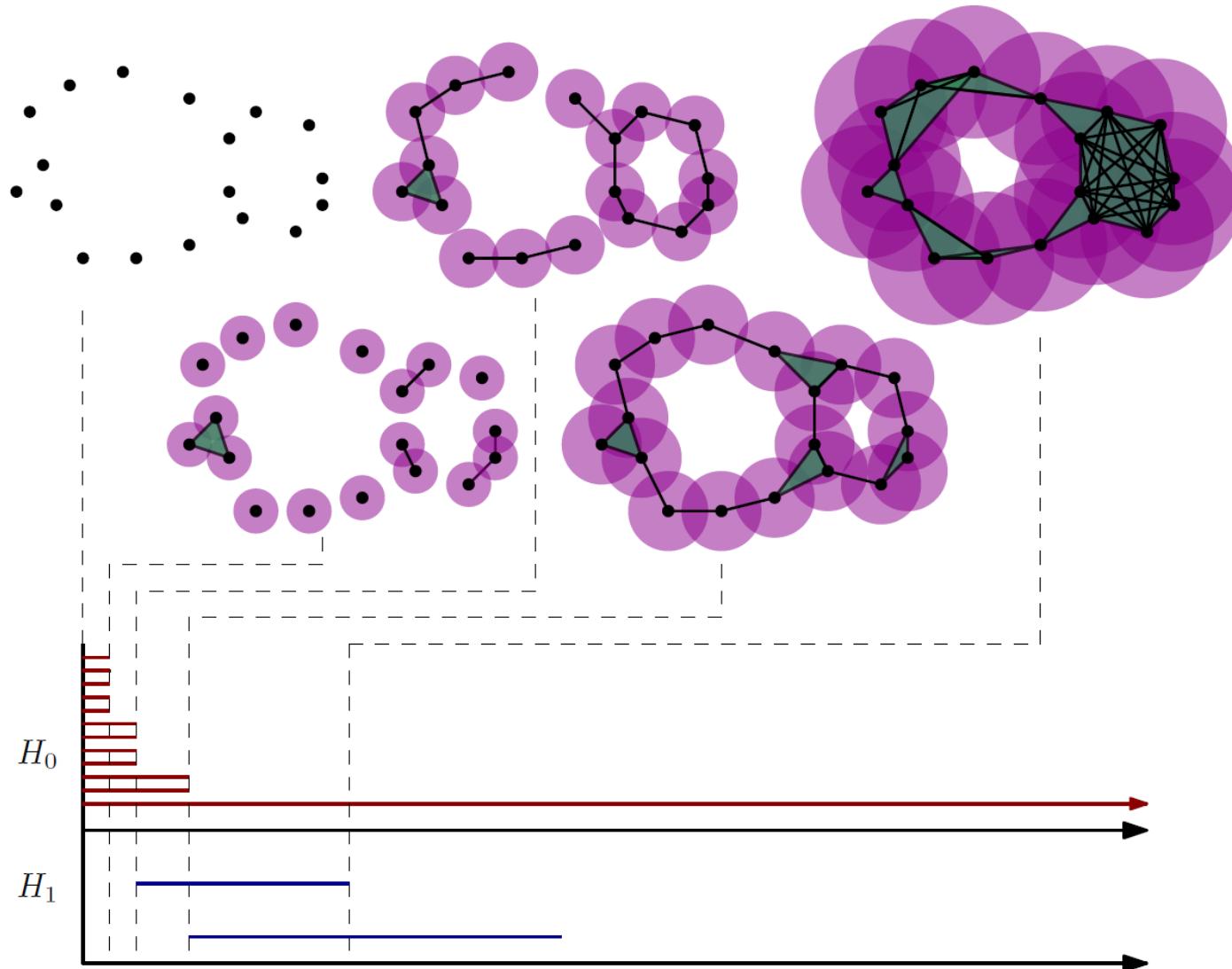
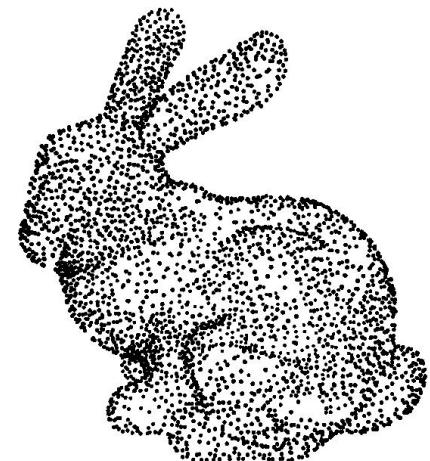


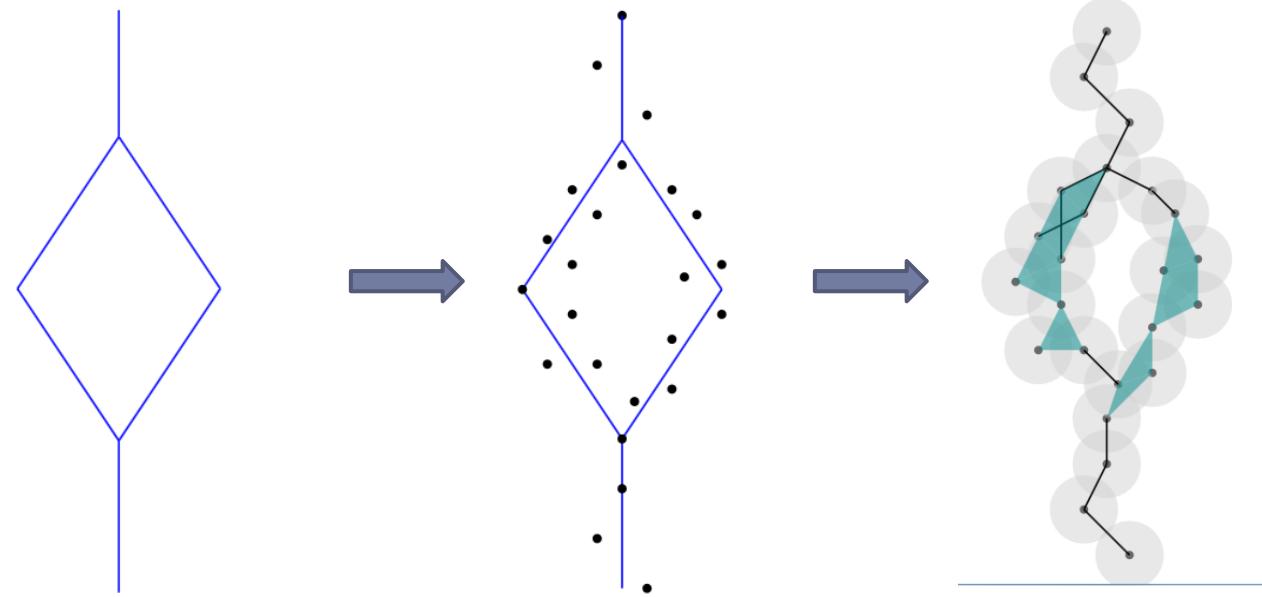
Image courtesy of T. K. Dey



Topological
summary
of hidden space

- ▶ Input:
 - ▶ A set of points $P \subseteq R^d$ sampled on/around X
- ▶ Question:
 - ▶ How well we approximate the **homology** of X
- ▶ We will be a little more precise on how well we can approximate these topological quantities today.

Pipeline



- ▶ PCD → simplicial complex → homology estimation
- ▶ What follows:
 - ▶ Sampling conditions
 - ▶ Inference guarantees

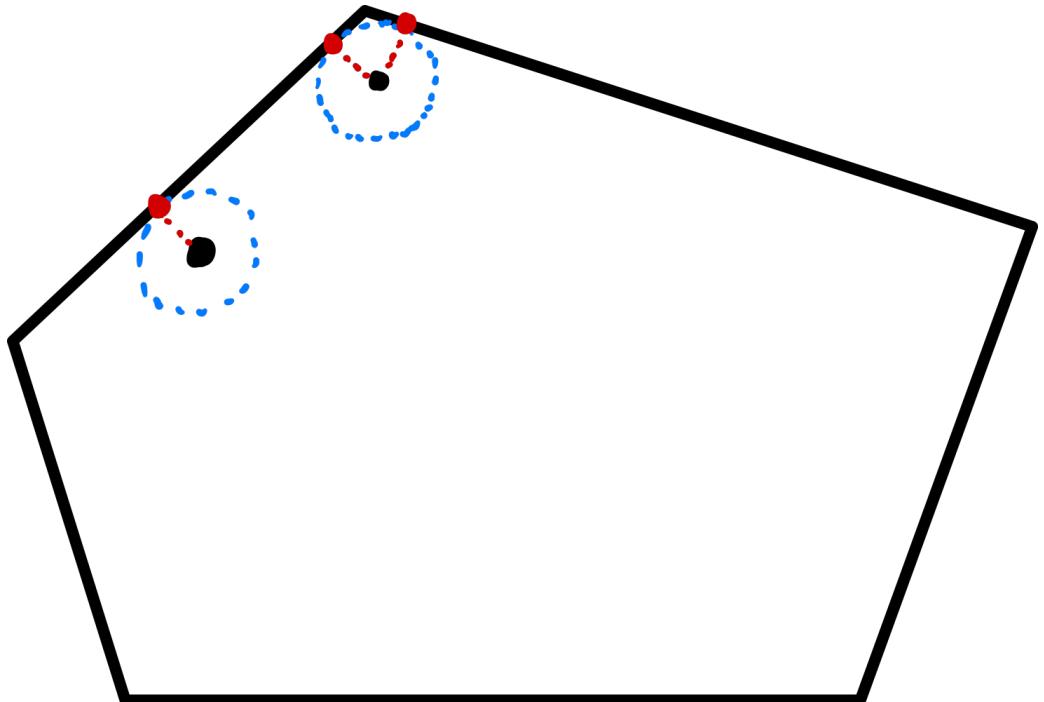
Sampling Conditions

Motivation

- ▶ Theoretical guarantees are usually obtained when input points P sampling the hidden domain “well enough”.
- ▶ Need to quantify the “wellness”.
- ▶ Two common ones based on:
 - ▶ Local feature size (for smooth manifolds)
 - ▶ Weak feature size (for general compact subspace of \mathbb{R}^d)

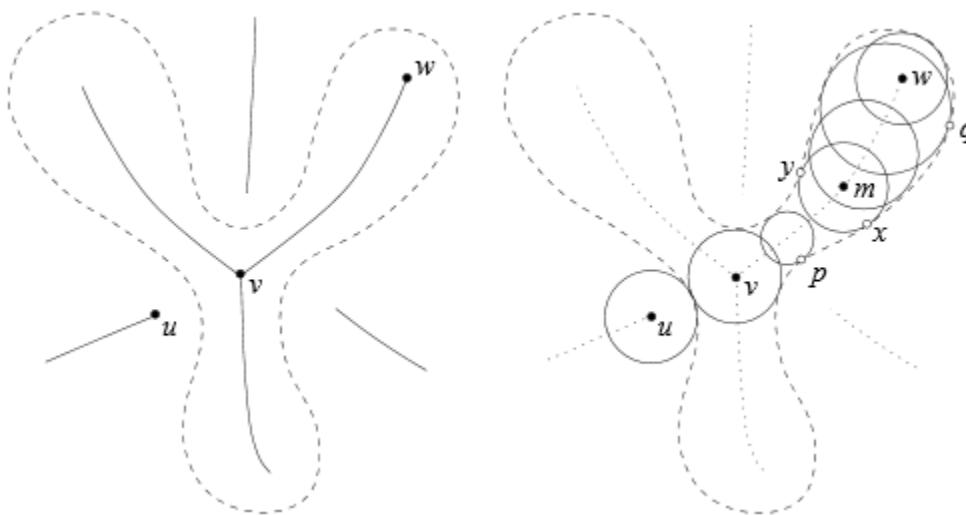
Distance Function

- ▶ $X \subset \mathbb{R}^d$: a compact subset of \mathbb{R}^d
- ▶ Distance function to set $d_X : \mathbb{R}^d \rightarrow [0, \infty)$
 - ▶ $d_X(x) = \min_{y \in X} d(x, y)$
 - ▶ d_X is a 1-Lipschitz function
- ▶ X^α : α -offset of X
 - ▶ $X^\alpha = \{y \in \mathbb{R}^d \mid d_X(y) \leq \alpha\}$
 - ▶ X^α is the sub level set $d_X^{-1}((-\infty, \alpha])$ of d_X
- ▶ Given any point $x \in \mathbb{R}^d$
 - ▶ $\Gamma(x) := \{ y \in X \mid d(x, y) = d_X(x) \}$



Medial Axis

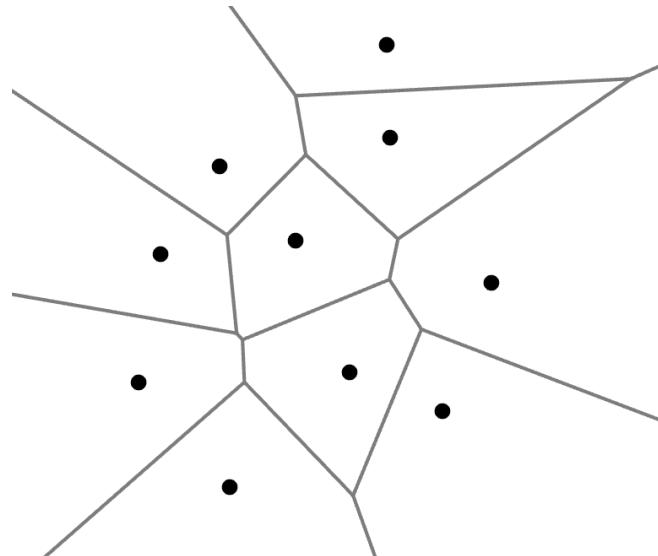
- ▶ The *medial axis* Σ of X is the closure of the set of points $x \in \mathbb{R}^d$ such that $|\Gamma(x)| \geq 2$
 - ▶ $|\Gamma(x)| \geq 2$ means that there is a medial ball $B_r(x)$ touching X at more than 1 point and whose interior is empty of points from X .



Courtesy of [Dey, 2006]

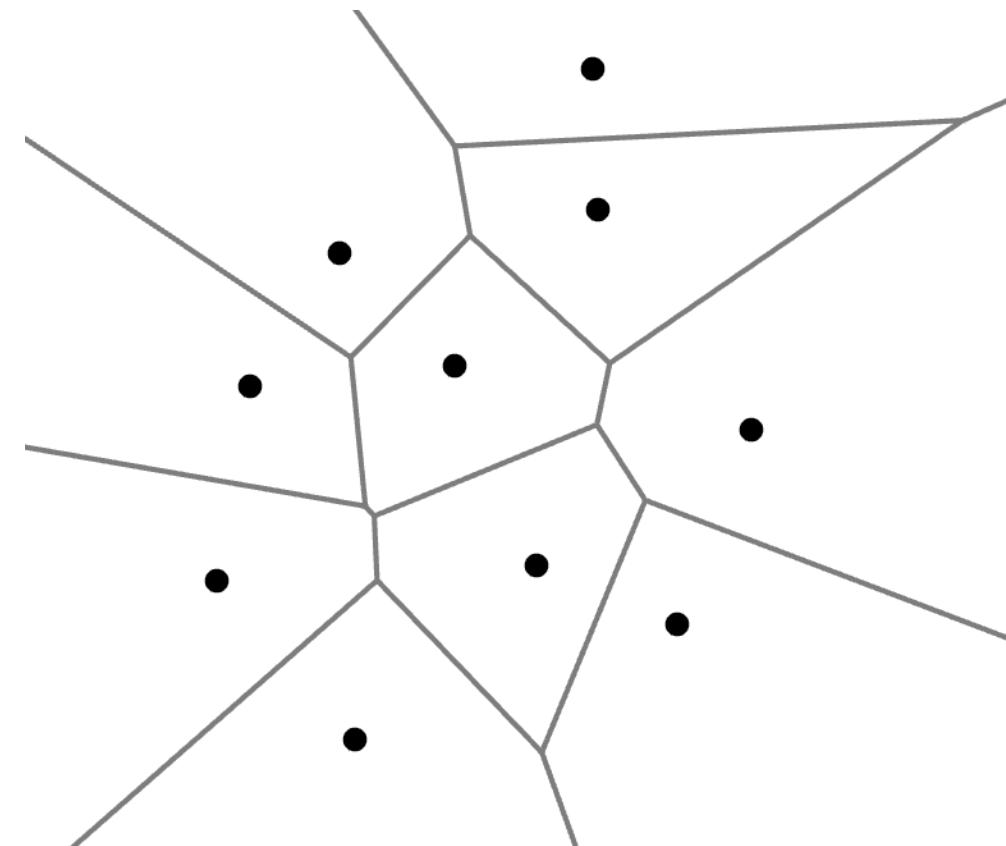
Voronoi Diagram

- Given a finite set $P = \{p_1, \dots, p_n\} \subset \mathbb{R}^d$, the **Voronoi cell** of p_i is
 - $Vor(p_i) = \{x \in \mathbb{R}^d \mid \|x - p_i\| \leq \|x - p_j\|, \forall j \neq i\}$
- The **Voronoi Diagram** of P is the collection of all Voronoi cells.



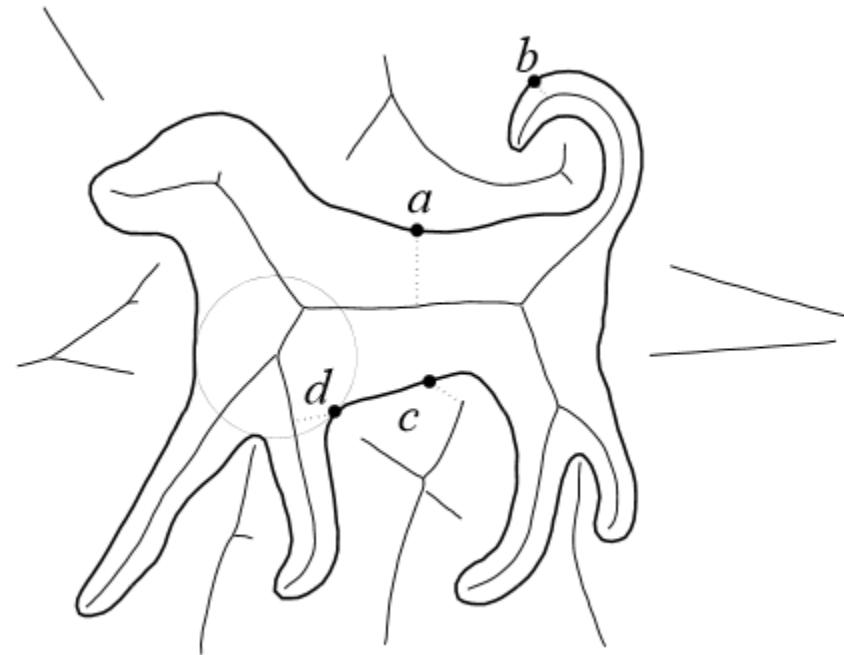
Medial Axis vs Voronoi Diagram

- When X is a finite set of points. Then, the medial axis agrees with the boundary of Voronoi cells.



Local Feature Size

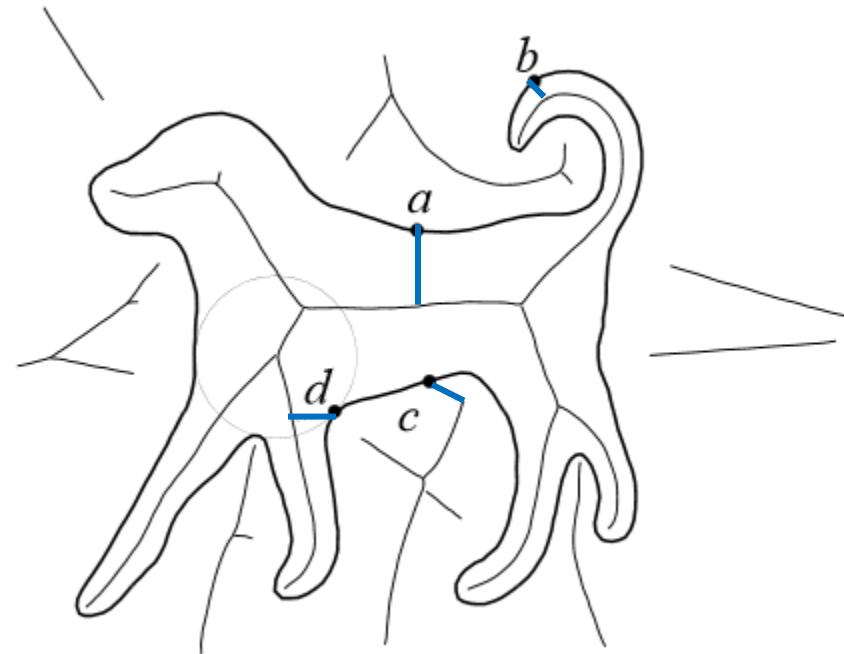
- ▶ The local feature size $lfs(x)$ at a point $x \in X$ is the distance of x to the medial axis Σ of X
 - ▶ That is, $lfs(x) = d(x, \Sigma)$
- ▶ This concept is adaptive
 - ▶ Large in a place without “features”
- ▶ Intuitively:
 - ▶ We should sample more densely if local feature size is small.
- ▶ The **reach** $\rho(X) = \inf_{x \in X} lfs(x)$



Courtesy of [Dey, 2006]

Local Feature Size

- ▶ The local feature size $lfs(x)$ at a point $x \in X$ is the distance of x to the medial axis Σ of X
 - ▶ That is, $lfs(x) = d(x, \Sigma)$
- ▶ This concept is adaptive
 - ▶ Large in a place without “features”
- ▶ Intuitively:
 - ▶ We should sample more densely if local feature size is small.
- ▶ The **reach** $\rho(X) = \inf_{x \in X} lfs(x)$



Courtesy of [Dey, 2006]

Smooth Manifold Case

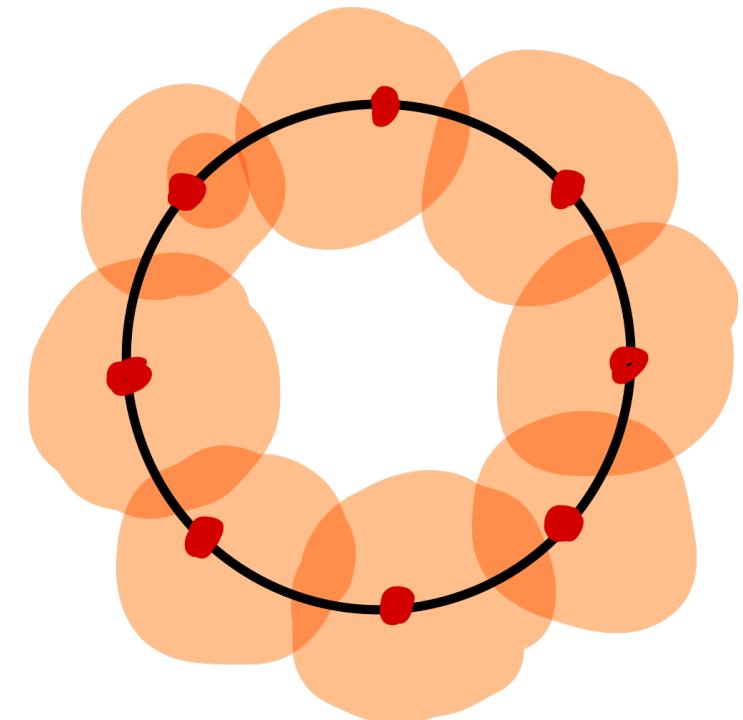
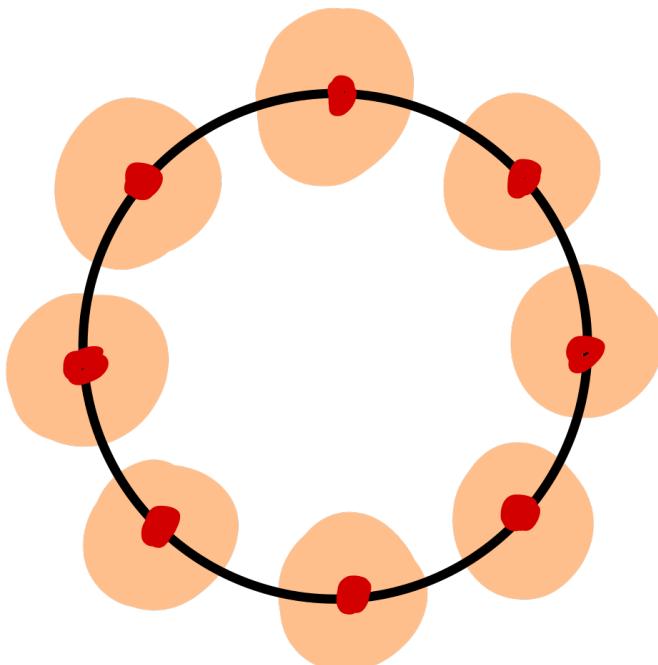
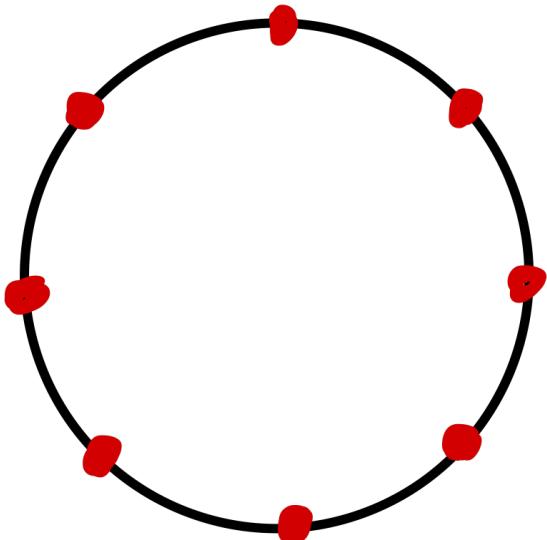
- Let X be a smooth manifold embedded in R^d

Theorem [Niyogi, Smale, Weinberger]

Let $P \subset X$ be such that $d_H(X, P) \leq \epsilon$. If $2\epsilon \leq \alpha \leq \sqrt{\frac{3}{5}}\rho(X)$,
there is a deformation retraction from P^α to X .

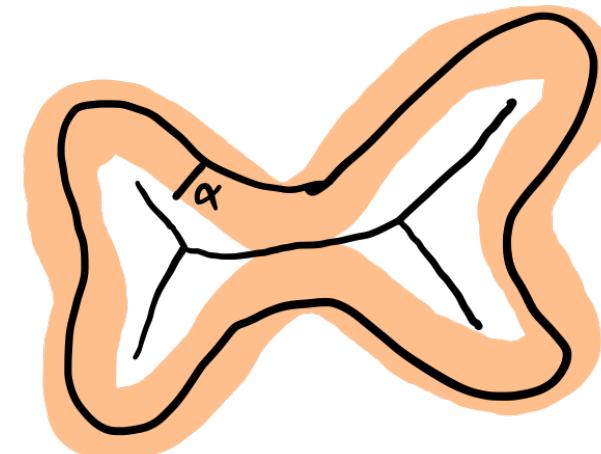
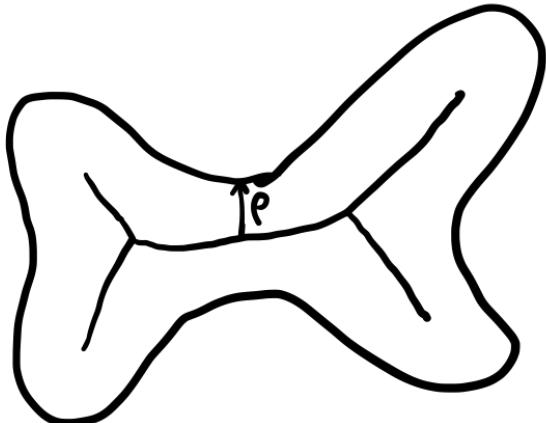
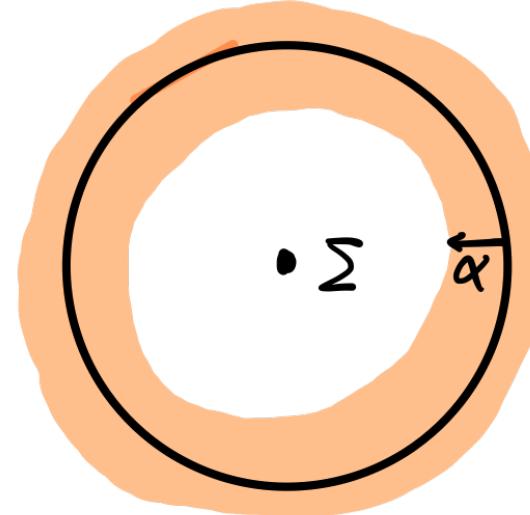
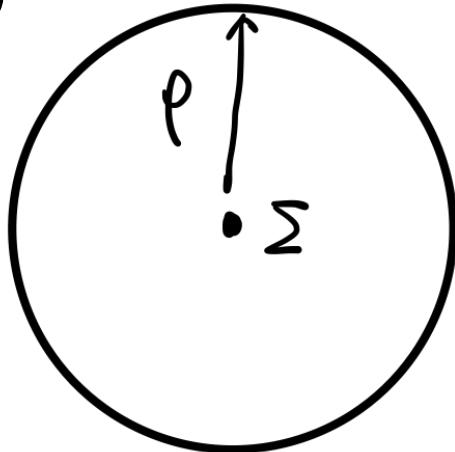
Interpretation

► $\alpha \geq 2\epsilon$

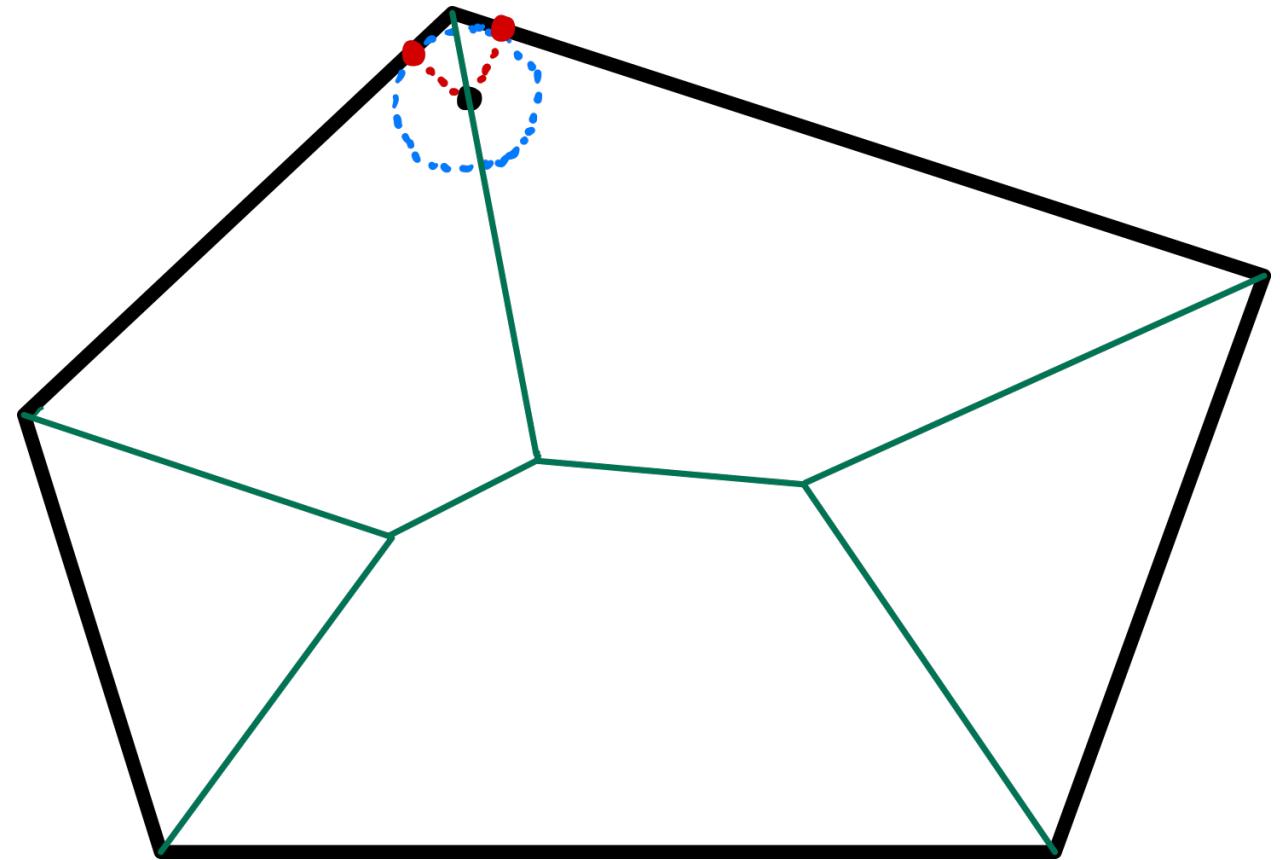


Interpretation

► $\alpha \leq \sqrt{\frac{3}{5}}\rho(X)$

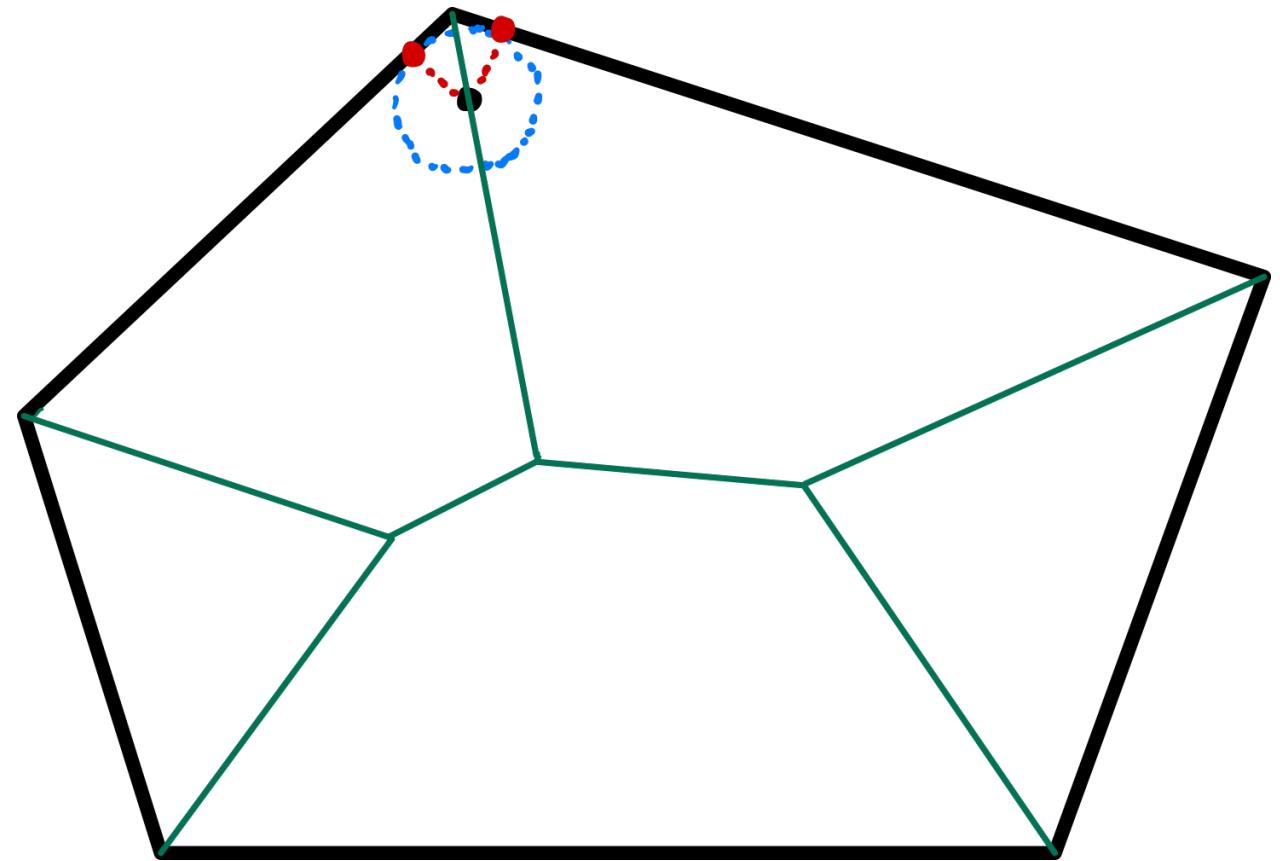


Non-smooth case



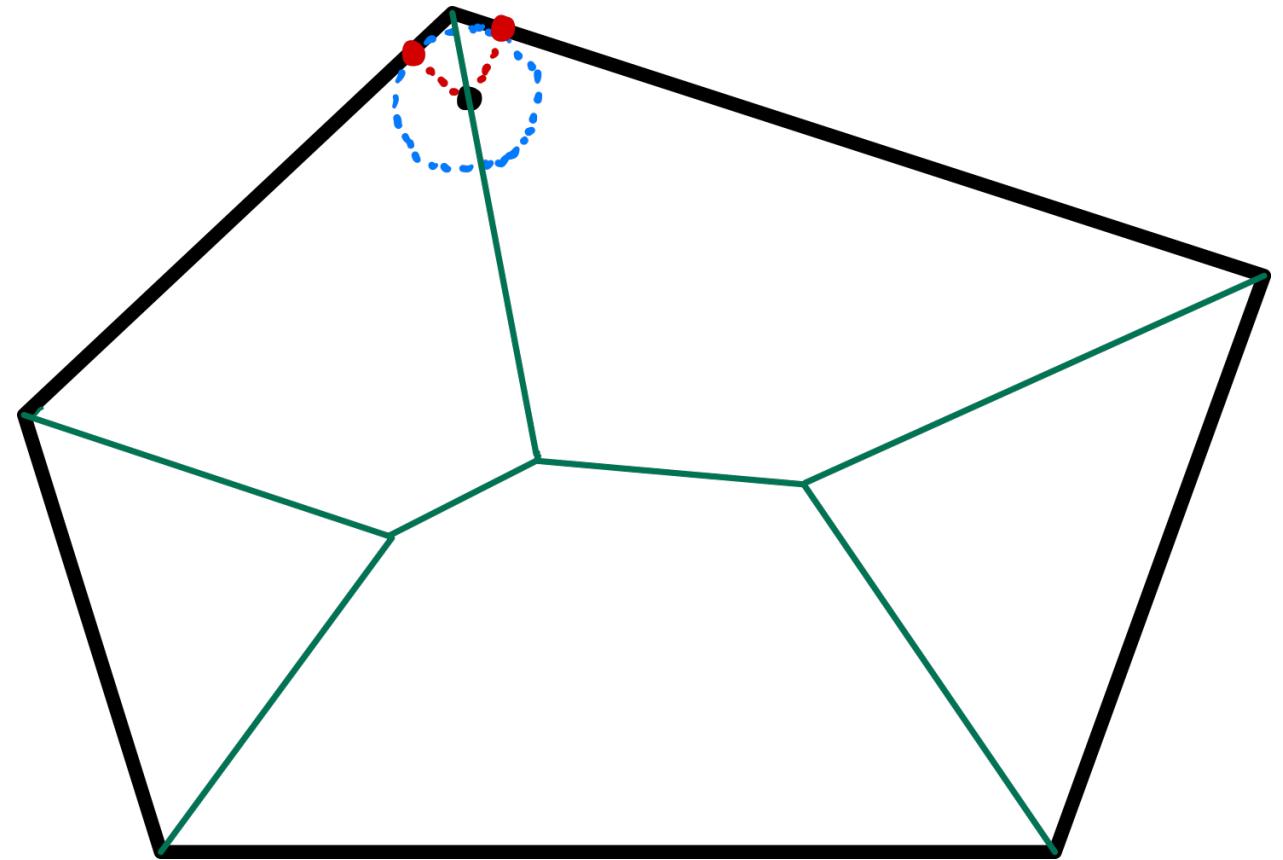
Non-smooth case

- ▶ $\rho(X)$ can be 0 for non-smooth shapes



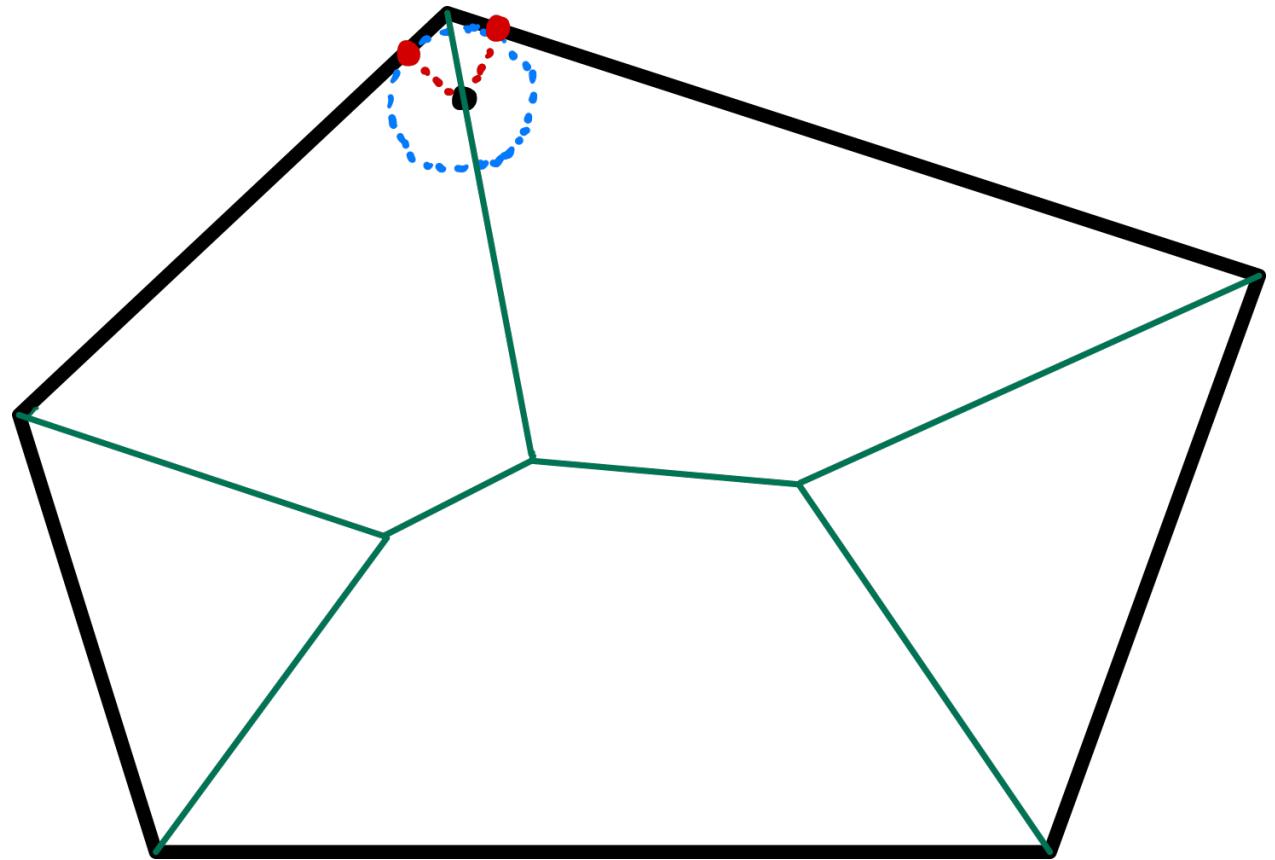
Non-smooth case

- ▶ $\rho(X)$ can be 0 for non-smooth shapes
- ▶ It is then not suitable for general compact spaces



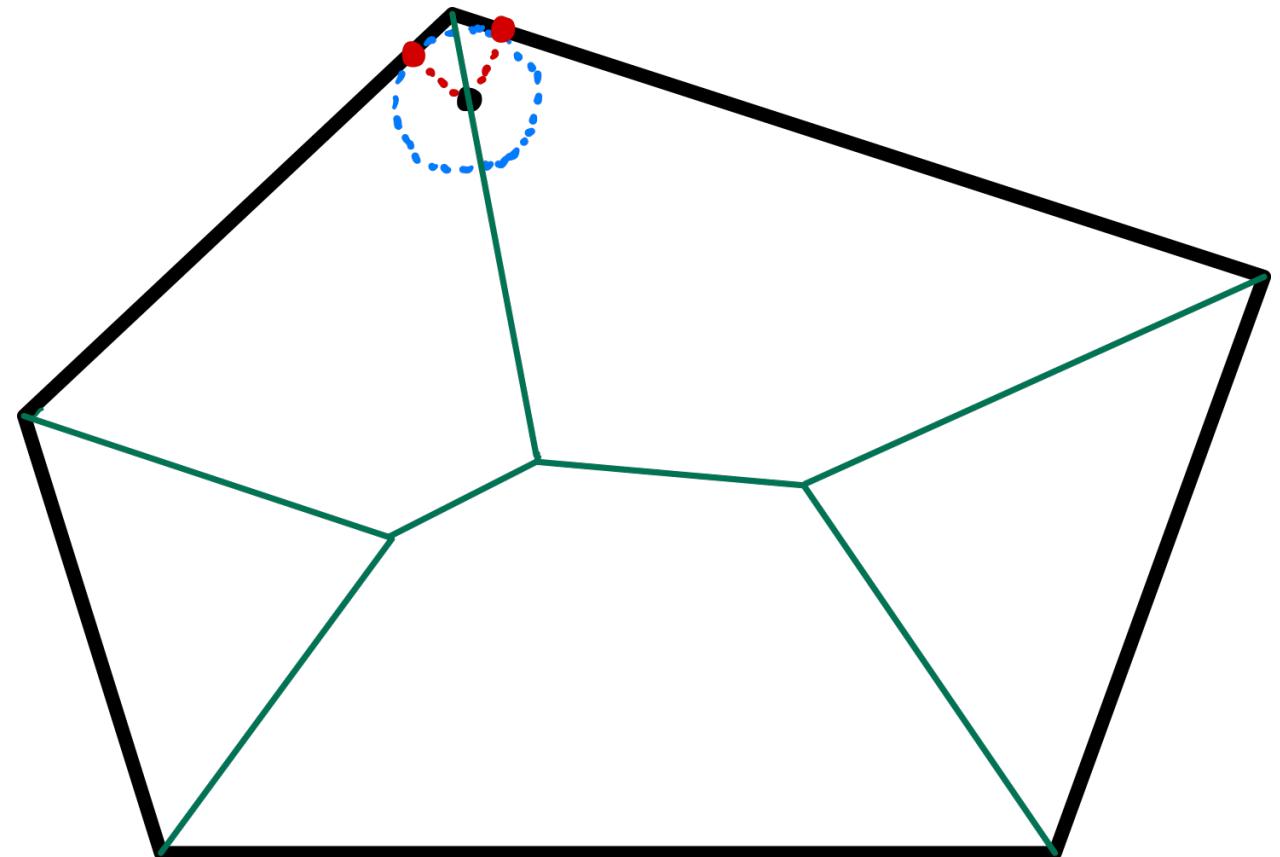
Non-smooth case

- ▶ $\rho(X)$ can be 0 for non-smooth shapes
- ▶ It is then not suitable for general compact spaces
- ▶ We are interested in topological changes of $d_X^{-1}((-\infty, \alpha])$



Non-smooth case

- ▶ $\rho(X)$ can be 0 for non-smooth shapes
- ▶ It is then not suitable for general compact spaces
- ▶ We are interested in topological changes of $d_X^{-1}((-\infty, \alpha])$
- ▶ This reminds us about Morse theory



Recall: Morse theory

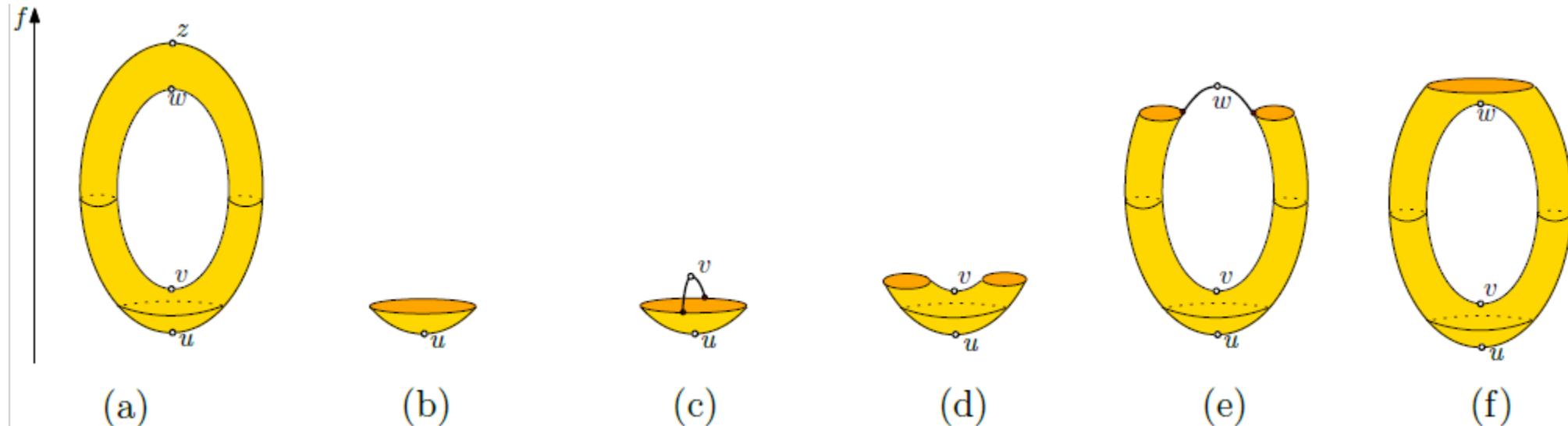
Theorem 3 (Homotopy type of sub-level sets). *Let $f : M \rightarrow \mathbb{R}$ be a smooth function defined on a manifold M . Given $a < b$, suppose the interval-level set $M_{[a,b]} = f^{-1}([a,b])$ is compact and contains no critical points of f . Then $M_{\leq a}$ is diffeomorphic to $M_{\leq b}$.*

Furthermore, $M_{\leq a}$ is a deformation retract of $M_{\leq b}$, and the inclusion map $i : M_{\leq a} \hookrightarrow M_{\leq b}$ is a homotopy equivalence.

Recall: Morse theory

Theorem 3 (Homotopy type of sub-level sets). *Let $f : M \rightarrow \mathbb{R}$ be a smooth function defined on a manifold M . Given $a < b$, suppose the interval-level set $M_{[a,b]} = f^{-1}([a,b])$ is compact and contains no critical points of f . Then $M_{\leq a}$ is diffeomorphic to $M_{\leq b}$.*

Furthermore, $M_{\leq a}$ is a deformation retract of $M_{\leq b}$, and the inclusion map $i : M_{\leq a} \hookrightarrow M_{\leq b}$ is a homotopy equivalence.



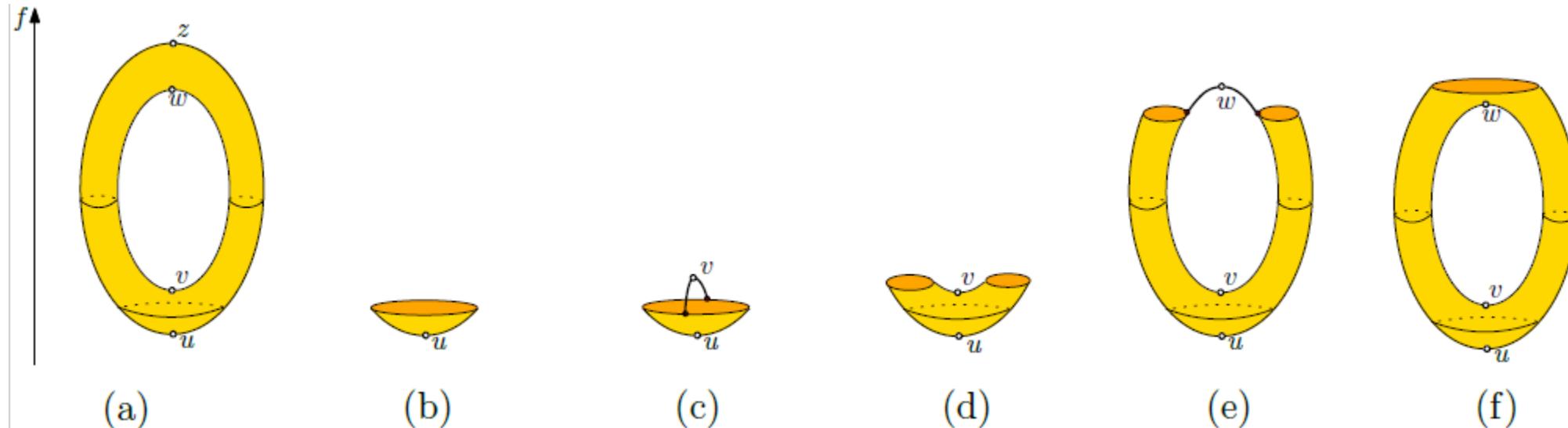
Recall: Morse theory

Theorem 4. *Given a Morse function $f : M \rightarrow \mathbb{R}$ defined on a smooth manifold M , let p be an index- k critical point of f with $\alpha = f(p)$. Assume $f^{-1}([\alpha - \varepsilon, \alpha + \varepsilon])$ is compact for a sufficiently small $\varepsilon > 0$ such that there is no other critical points of f contained in this interval-level set other than p . Then the sublevel set $M_{\leq \alpha+\varepsilon}$ has the same homotopy type as $M_{\leq \alpha-\varepsilon}$ with a k -cell attached to its boundary $\text{Bd } M_{\leq \alpha-\varepsilon}$.*

Animation

Recall: Morse theory

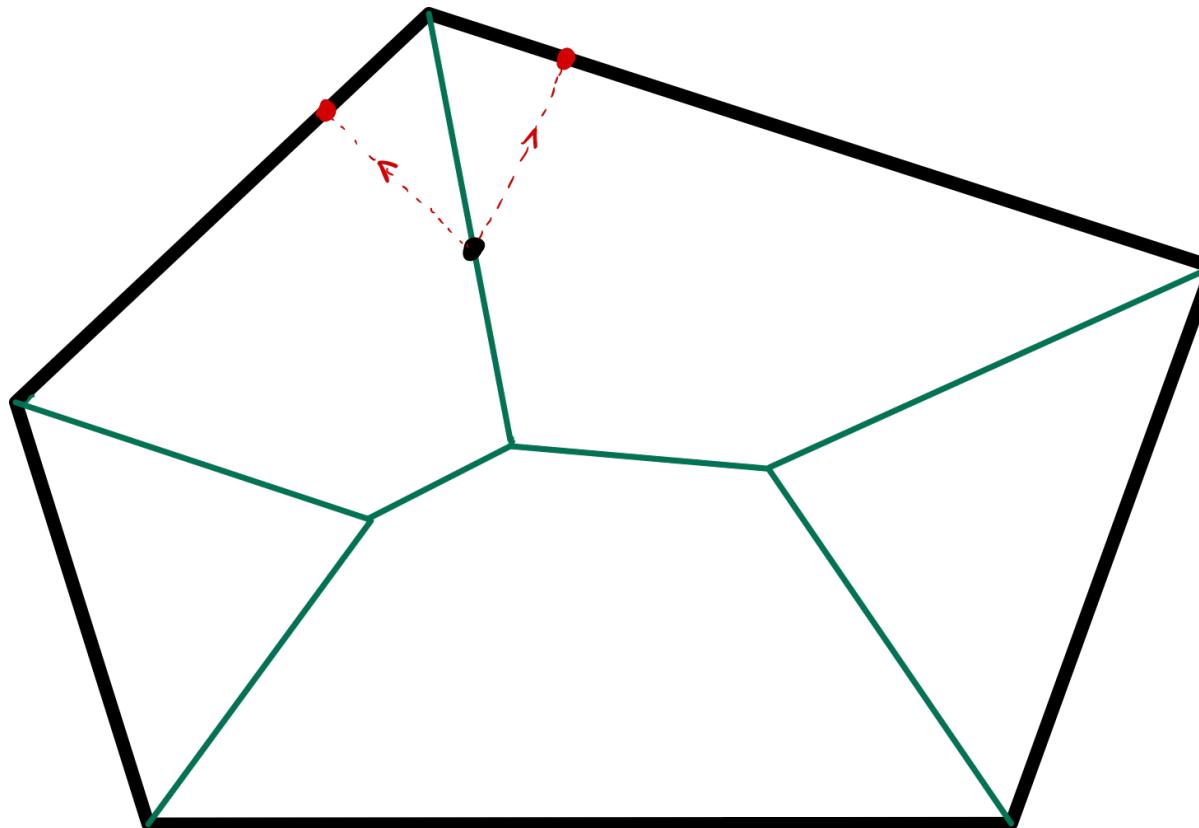
Theorem 4. Given a Morse function $f : M \rightarrow \mathbb{R}$ defined on a smooth manifold M , let p be an index- k critical point of f with $\alpha = f(p)$. Assume $f^{-1}([\alpha - \varepsilon, \alpha + \varepsilon])$ is compact for a sufficiently small $\varepsilon > 0$ such that there is no other critical points of f contained in this interval-level set other than p . Then the sublevel set $M_{\leq \alpha + \varepsilon}$ has the same homotopy type as $M_{\leq \alpha - \varepsilon}$ with a k -cell attached to its boundary $\text{Bd } M_{\leq \alpha - \varepsilon}$.



Animation

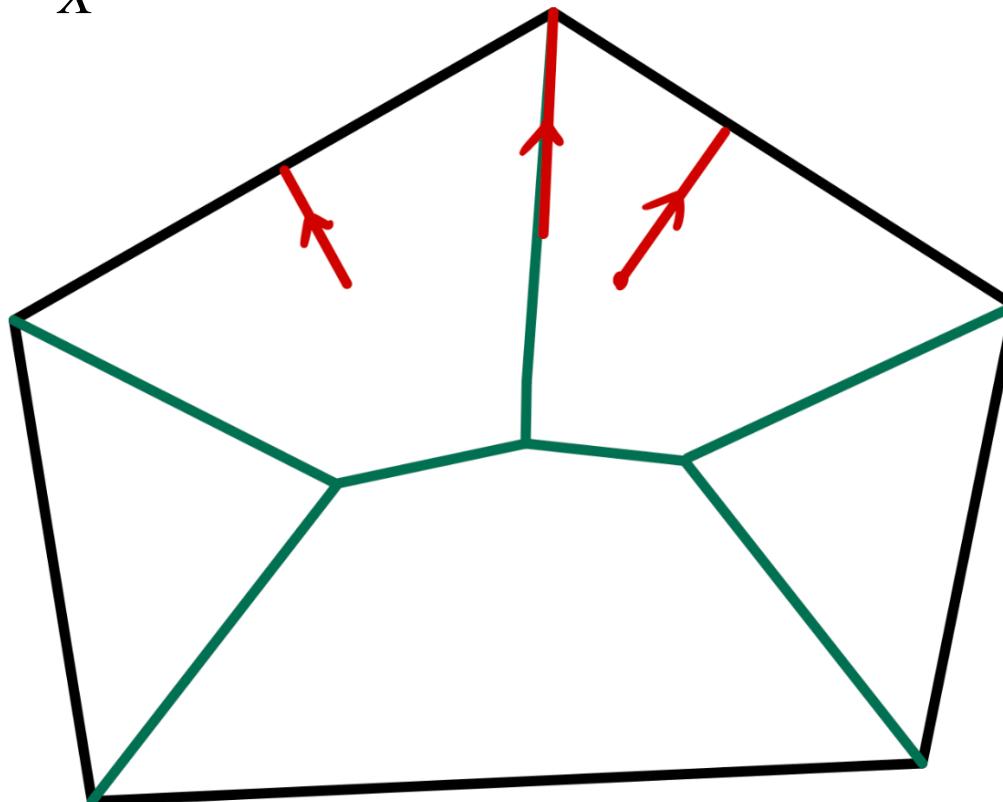
Morse theory for distance function?

- ▶ Distance function d_X not differentiable on $\Sigma \cup X$



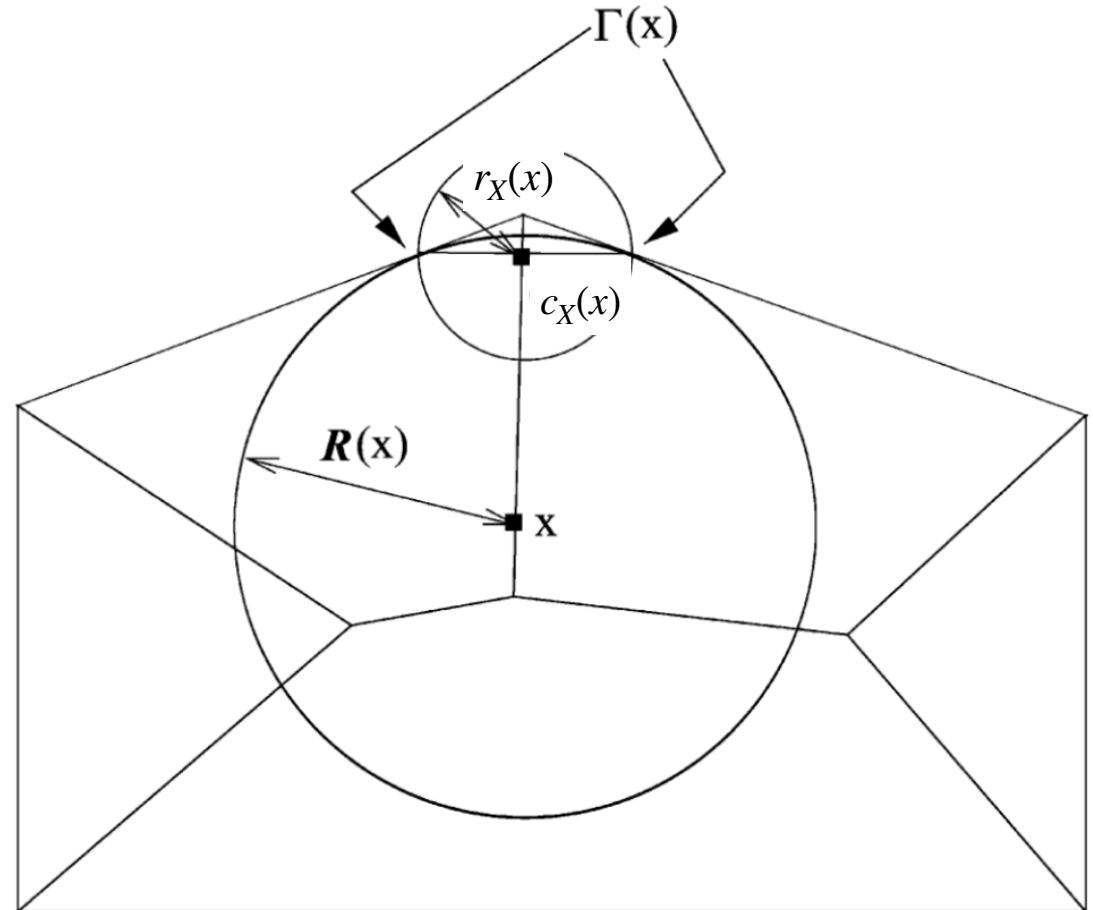
Gradient of Distance Function

- ▶ Can we take the average of different directions for points in Σ to define gradient of d_X ?



Gradient of Distance Function

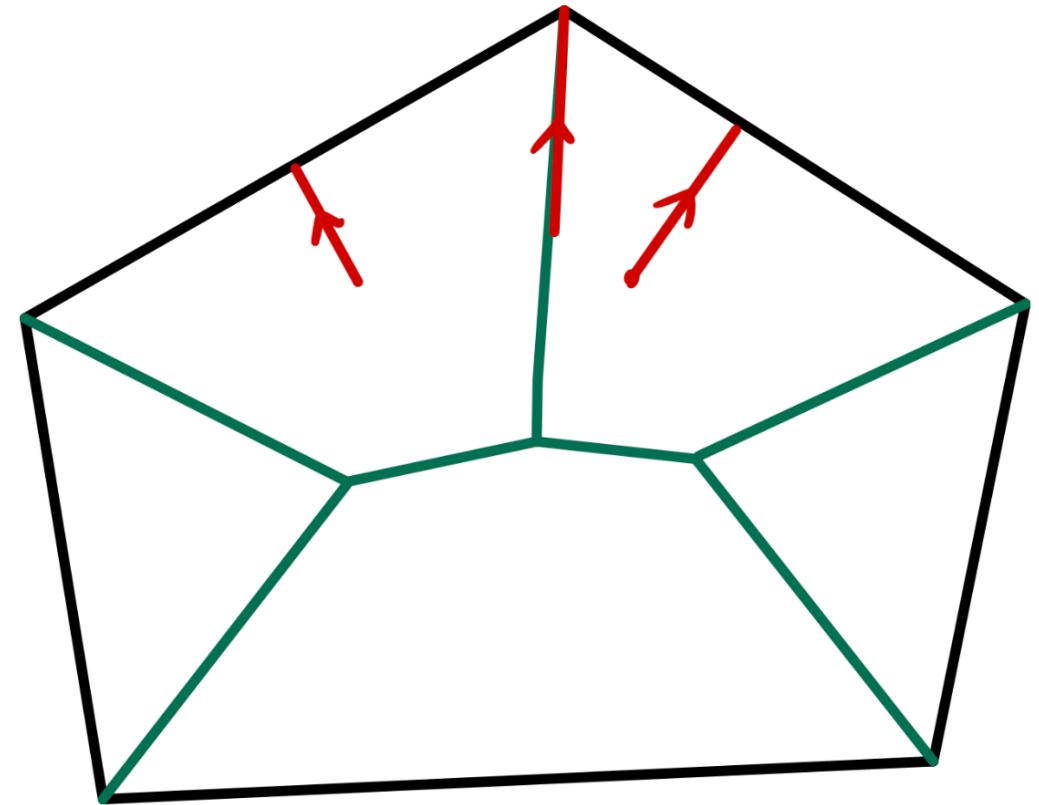
- ▶ Distance function not differentiable on the $\Sigma \cup X$
- ▶ Still can define a generalized concept of gradient
 - ▶ [Grove '93] [Lieutier, 2004]
- ▶ For $x \in \mathbb{R}^d \setminus X$,
 - ▶ Let $c_X(x)$ and $r_X(x)$ be the center and radius of the smallest enclosing ball of point(s) in $\Gamma(x)$
 - ▶ The *generalized gradient* of distance function
- $$\nabla_X(x) = \frac{x - c_X(x)}{d_X(x)}$$
- ▶ Flow lines induced by the generalized gradient



Courtesy of Lieutier 2004

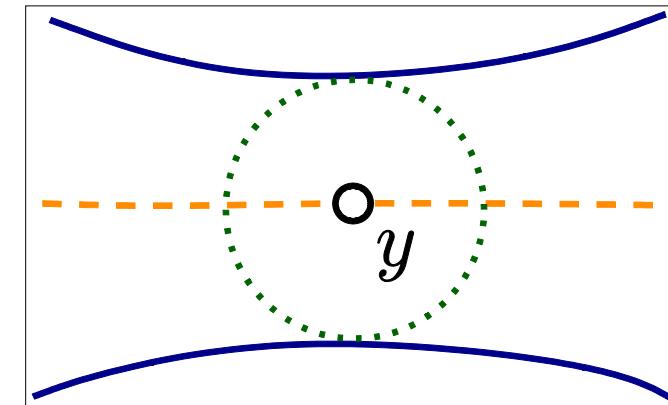
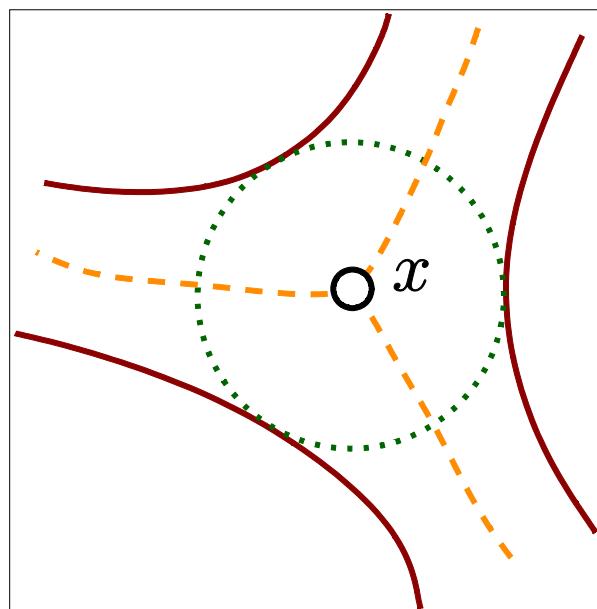
Gradient of Distance Function

- ▶ Distance function not differentiable on the $\Sigma \cup X$
- ▶ Still can define a generalized concept of gradient
 - ▶ [Grove '93] [Lieutier, 2004]
- ▶ For $x \in \mathbb{R}^d \setminus X$,
 - ▶ Let $c_X(x)$ and $r_X(x)$ be the center and radius of the smallest enclosing ball of point(s) in $\Gamma(x)$
 - ▶ The *generalized gradient* of distance function
- $$\nabla_X(x) = \frac{x - c_X(x)}{d_X(x)}$$
- ▶ Flow lines induced by the generalized gradient



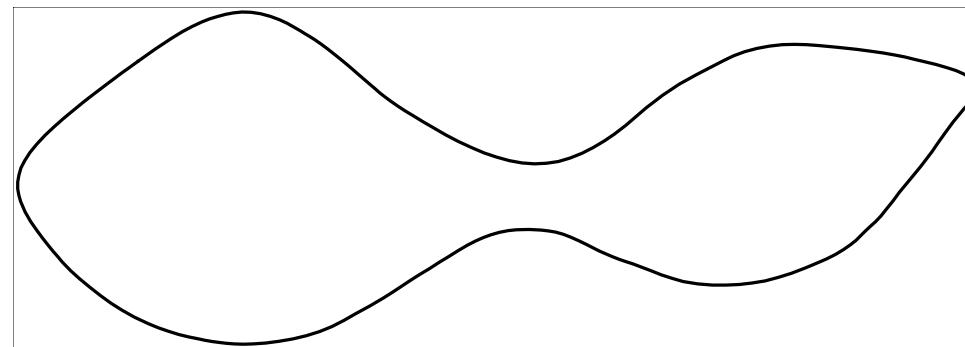
Critical Points

- ▶ A critical point of the distance function is a point whose generalized gradient $\nabla_X(x)$ vanishes, i.e., $c_X(x)$ coincides with x
- ▶ A critical point is in its medial axis Σ



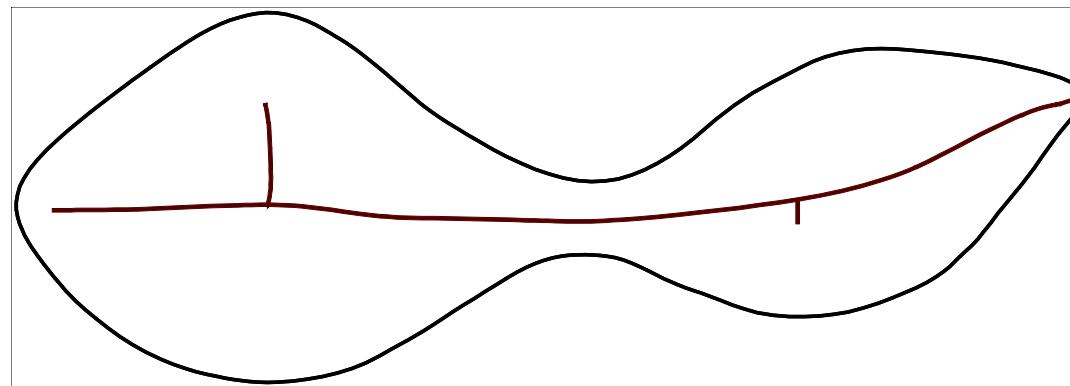
Critical Points

- ▶ A critical point of the distance function is a point whose generalized gradient $\nabla_X(x)$ vanishes
- ▶ A critical point is in its medial axis Σ



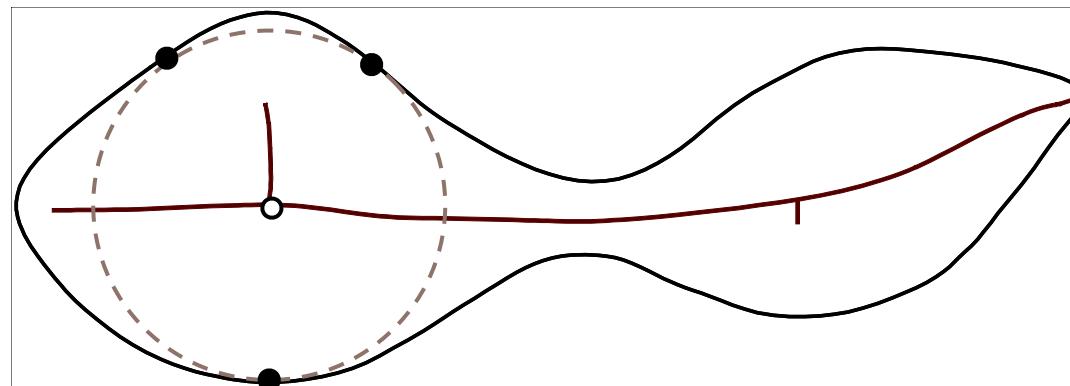
Critical Points

- ▶ A critical point of the distance function is a point whose generalized gradient $\nabla_X(x)$ vanishes
- ▶ A critical point is in its medial axis Σ



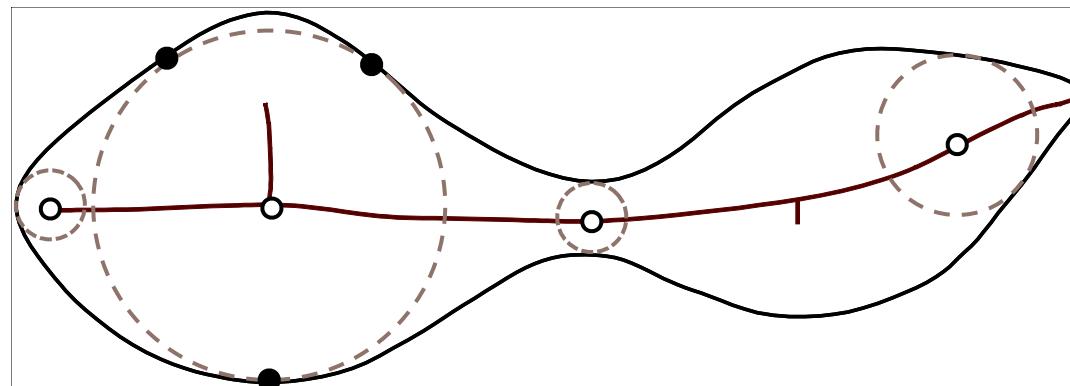
Critical Points

- ▶ A critical point of the distance function is a point whose generalized gradient $\nabla_X(x)$ vanishes
- ▶ A critical point is in its medial axis Σ



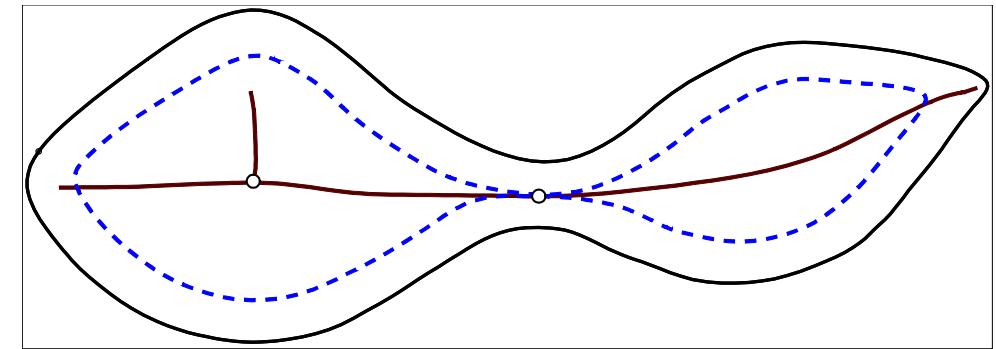
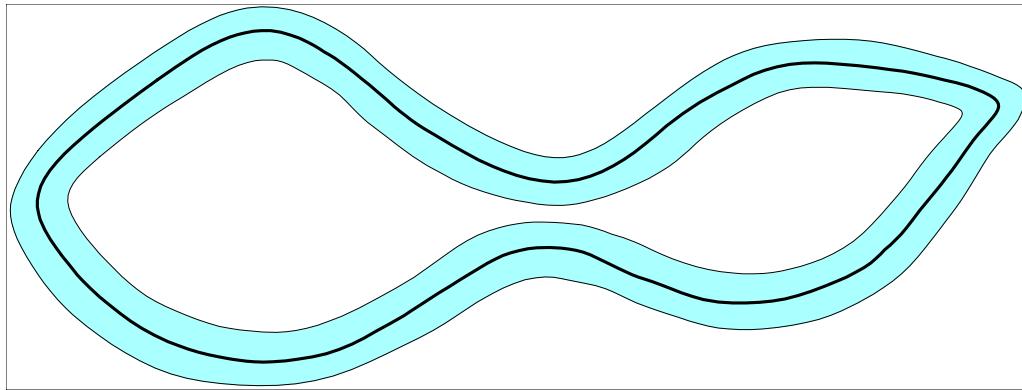
Critical Points

- ▶ A critical point of the distance function is a point whose generalized gradient $\nabla_X(x)$ vanishes
- ▶ A critical point is in its medial axis Σ



When does offset changes its topology?

- ▶ When they sweep past critical points!



Critical point theory for distance function

- ▶ **Theorem [Offset Homotopy] [Grove '93]**

If $0 < \alpha < \alpha'$ are such that there is no critical value of d_X in the closed interval $[\alpha, \alpha']$, then $X^{\alpha'}$ deformation retracts onto X^α . In particular, $H(X^\alpha) \cong H(X^{\alpha'})$.

- ▶ **Remarks:**

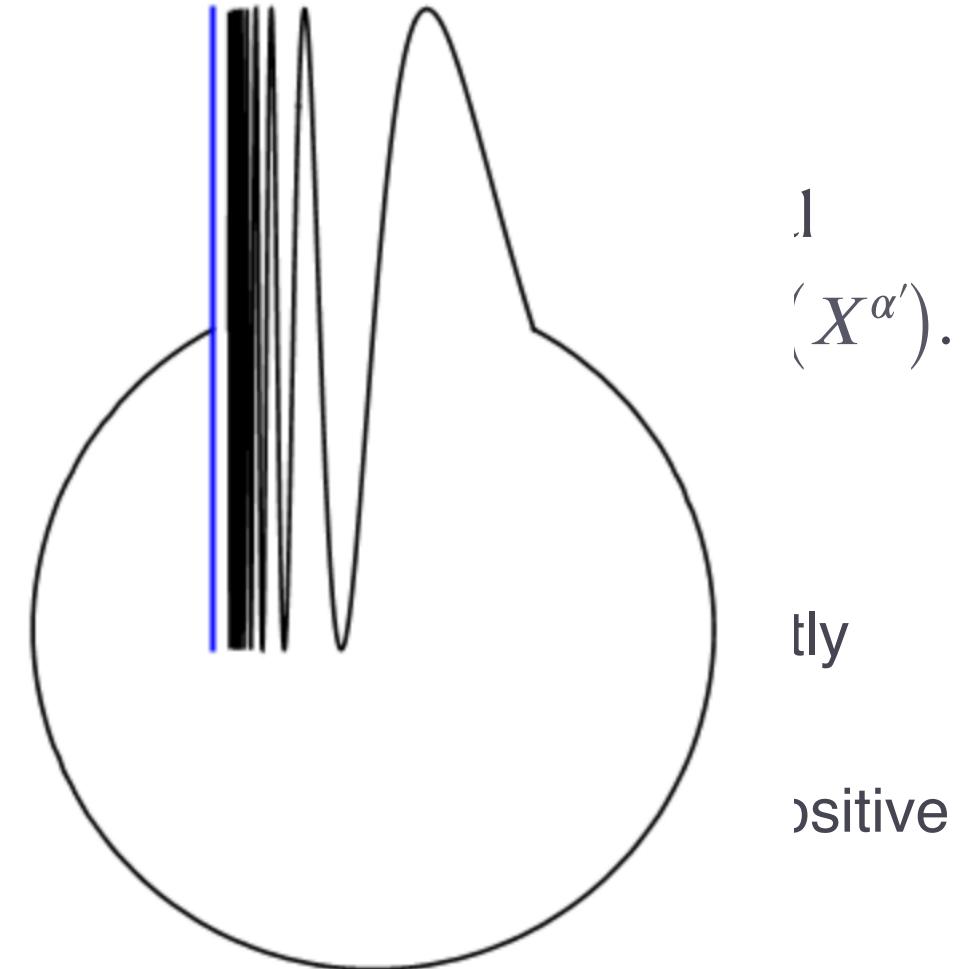
- ▶ For the case of compact set X , note that it is possible that X^α , for sufficiently small $\alpha > 0$, may not be homotopy equivalent to $X^0 = X$.
- ▶ Intuitively, by above theorem, we can approximate $H(X^\alpha)$ for any small positive α from a thickened version (offset) of X^α .

Critical point theory for distance function

- ▶ Theorem [Offset Homotopy] [*Grove'93*]

If $0 < \alpha < \alpha'$ are such that there is no critical v

[α, α'], then $X^{\alpha'}$ deformation retracts onto X



- ▶ Remarks:

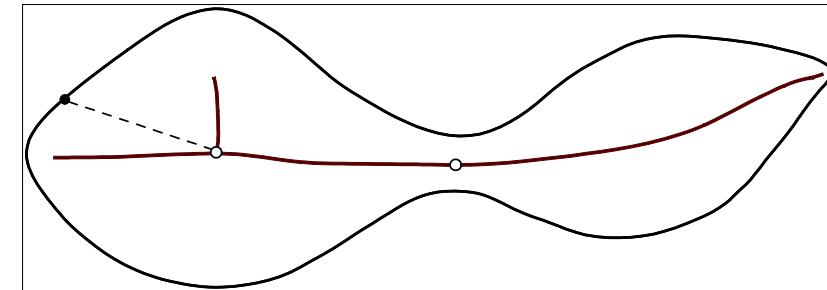
- ▶ For the case of compact set X , note that it is small $\alpha > 0$, may not be homotopy equivale
- ▶ Intuitively, by above theorem, we can approx α from a thickened version (offset) of X^α .

Weak Feature Size

- ▶ Given a compact $X \subset \mathbb{R}^d$, let $C \subset \mathbb{R}^d$ denote
 - ▶ the set of critical points of the distance function d_X
 - ▶ Note $C \subset \Sigma$
- ▶ Given a compact $X \subset \mathbb{R}^d$, the *weak feature size* is

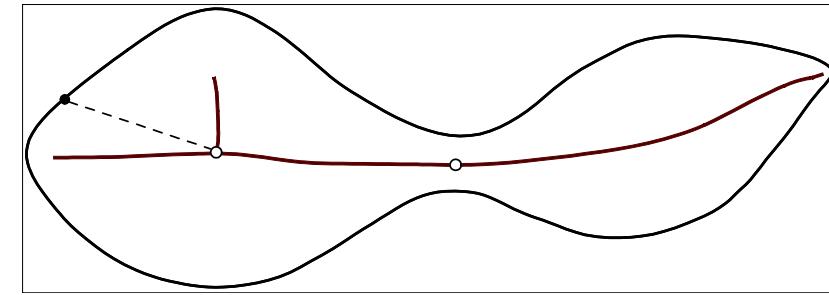
Weak Feature Size

- ▶ Given a compact $X \subset \mathbb{R}^d$, let $C \subset \mathbb{R}^d$ denote
 - ▶ the set of critical points of the distance function d_X
 - ▶ Note $C \subset \Sigma$
- ▶ Given a compact $X \subset \mathbb{R}^d$, the *weak feature size* is



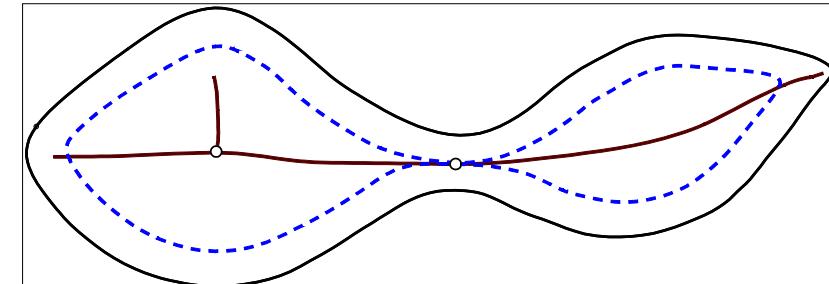
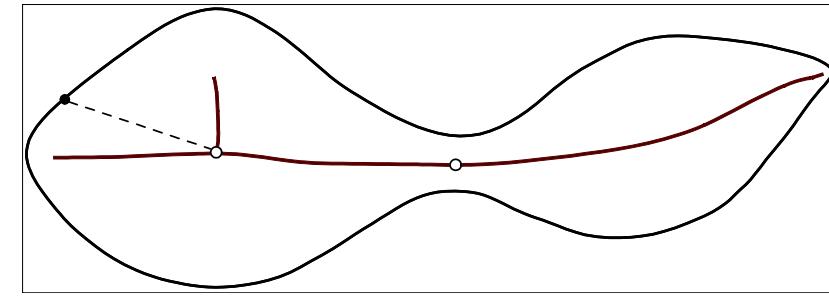
Weak Feature Size

- ▶ Given a compact $X \subset \mathbb{R}^d$, let $C \subset \mathbb{R}^d$ denote
 - ▶ the set of critical points of the distance function d_X
 - ▶ Note $C \subset \Sigma$
- ▶ Given a compact $X \subset \mathbb{R}^d$, the *weak feature size* is
 - ▶ $wfs(X) = \inf_{x \in X} d(x, C)$
- ▶ Equivalently,
 - ▶ $wfs(X)$ is the infimum of the positive critical value of d_X
- ▶ $\rho(X) \leq wfs(X)$

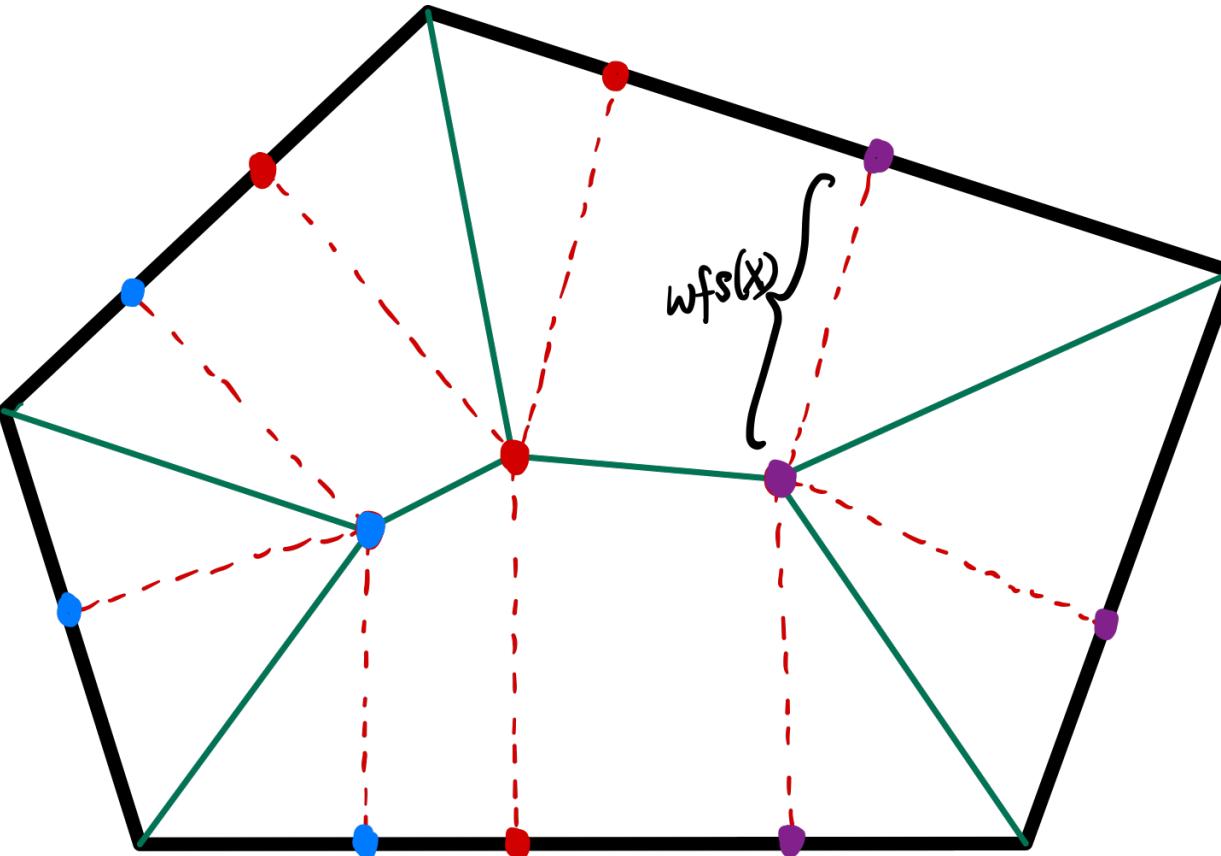


Weak Feature Size

- ▶ Given a compact $X \subset \mathbb{R}^d$, let $C \subset \mathbb{R}^d$ denote
 - ▶ the set of critical points of the distance function d_X
 - ▶ Note $C \subset \Sigma$
- ▶ Given a compact $X \subset \mathbb{R}^d$, the *weak feature size* is
 - ▶ $wfs(X) = \inf_{x \in X} d(x, C)$
- ▶ Equivalently,
 - ▶ $wfs(X)$ is the infimum of the positive critical value of d_X
 - ▶ $\rho(X) \leq wfs(X)$



Example

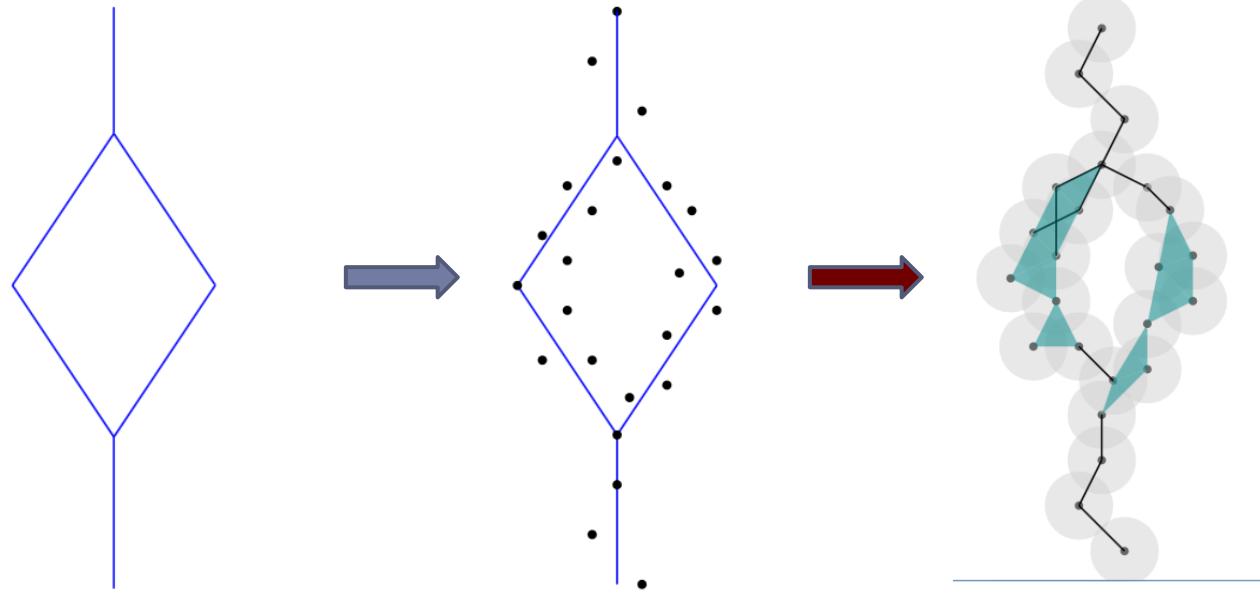


Typical Sampling Conditions

- ▶ Hausdorff distance $d_H(A, B)$ between two sets A and B
 - ▶ smallest value α such that $A \subseteq B^\alpha$ and $B \subseteq A^\alpha$
- ▶ No noise version:
 - ▶ A set of points P is an ϵ -sample of X
if $P \subset X$ and $d_H(P, X) \leq \epsilon$
- ▶ With noise version:
 - ▶ A set of points P is an ϵ -sample of X if $d_H(P, X) \leq \epsilon$
- ▶ The Hausdorff distance should be **smaller** than the reach (for manifold) or the weak feature size for general compact subsets

Homology Inference from PCD

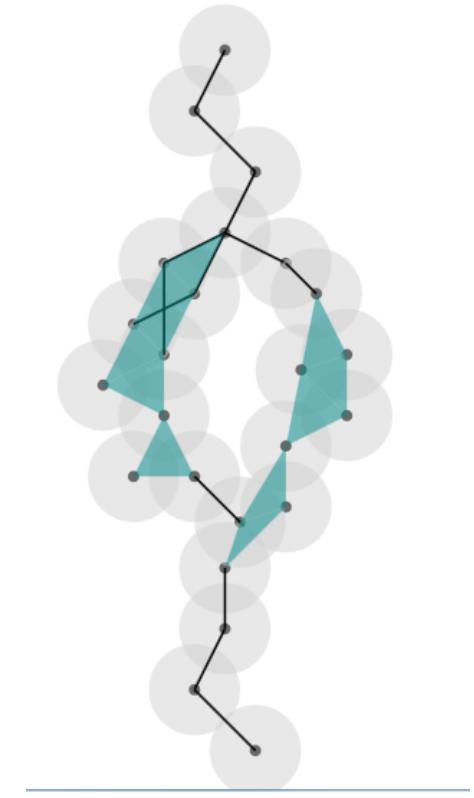
Problem Setup



- ▶ A hidden compact X (or a manifold M)
- ▶ An ϵ -sample P of X
- ▶ Recover homology of X from some complex built on P
 - ▶ will focus on Čech complex and Rips complex

Union of Balls

- ▶ $X^\alpha = \bigcup_{x \in X} B(x, \alpha)$
- ▶ $P^\alpha = \bigcup_{p \in P} B(p, \alpha)$
- ▶ Intuitively, P^α approximates offset X^α
- ▶ The Čech complex $C^\alpha(P)$ is the Nerve of P^α
- ▶ By Nerve Lemma, $C^\alpha(P)$ is homotopy equivalent to P^α



Smooth Manifold Case

- Let X be a smooth manifold embedded in \mathbb{R}^d

Theorem [Niyogi, Smale, Weinberger]

Let $P \subset X$ be such that $d_H(X, P) \leq \epsilon$. If $2\epsilon \leq \alpha \leq \sqrt{\frac{3}{5}}\rho(X)$,
there is a deformation retraction from P^α to X .

Corollary A

Under the conditions above, we have

$$H_*(X) \cong H_*(P^\alpha)$$

Convert to Čech Complexes

- ▶ Lemma A [*Chazal and Oudot, 2008*]:
- ▶ The following diagram commutes:

$$\begin{array}{ccc} H(P^\alpha) & \xrightarrow{i_*} & H(P^\beta) \\ h_* \downarrow & & \downarrow h_* \\ H(C^\alpha) & \xrightarrow{i_*} & H(C^\beta) \end{array}$$

Corollary B

Let $P \subset X$ be s.t. $d_H(X, P) \leq \epsilon$. If

$$2\epsilon \leq \alpha \leq \beta \leq \sqrt{\frac{3}{5}}\rho(X), H_*(X) \cong H_*(C^\alpha) \cong H_*(C^\beta)$$

where the second isomorphism is induced by inclusion.

- ▶ How about using Rips complex instead of Čech complex?
- ▶ Recall that

$$C^r(P) \subseteq R^r(P) \subseteq C^{2r}(P)$$

inducing

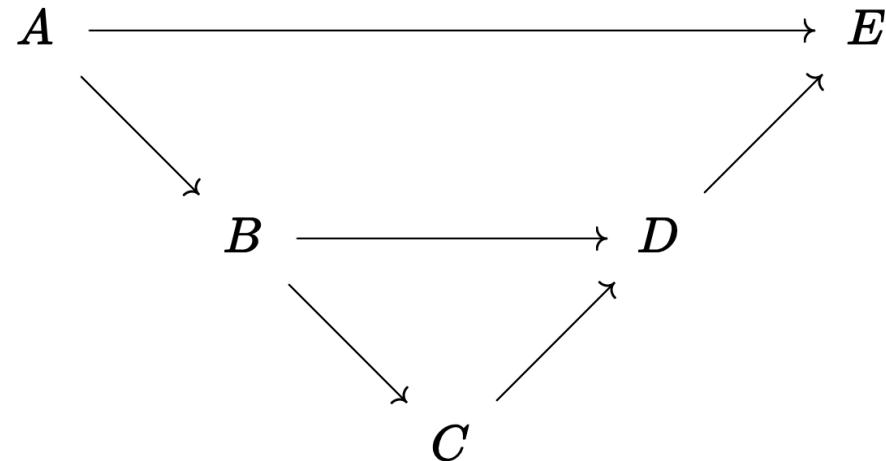
$$H(C^r) \rightarrow H(R^r) \rightarrow H(C^{2r})$$

- ▶ Idea [*Chazal and Oudot 2008*]:
 - ▶ Forming interleaving sequence of homomorphism to connect them with the homology of the input manifold X and its offsets X^α

From Rips Complex

- ▶ **Lemma B:**

Given a sequence $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$ of linear maps between finite dimensional vector spaces, if $\text{rank}(A \rightarrow E) = \dim C$ then $\text{rank}(B \rightarrow D) = \dim C$.



From Rips Complex

- ▶ **Lemma B:**

Given a sequence $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$ of linear maps between finite dimensional vector spaces, if $\text{rank}(A \rightarrow E) = \dim C$ then $\text{rank}(B \rightarrow D) = \dim C$.

- ▶ **Rips and Čech complexes:**

$$\begin{aligned} C^\alpha(P) &\subseteq R^\alpha(P) \subseteq C^{2\alpha}(P) \subseteq R^{2\alpha}(P) \subseteq C^{4\alpha}(P) \\ \Rightarrow H_*(C^\alpha) &\rightarrow H_*(R^\alpha) \rightarrow H_*(C^{2\alpha}) \rightarrow H_*(R^{2\alpha}) \rightarrow H_*(C^{4\alpha}) \end{aligned}$$

- ▶ **Applying Lemma B, if**
 $H_*(C^\alpha) \cong H_*(C^{2\alpha}) \cong H_*(C^{4\alpha}) \cong H_*(X)$ then

$$\text{rank}(H_*(R^\alpha) \rightarrow H_*(R^{2\alpha})) = \dim H_*(C^{2\alpha}) = \dim H_*(X)$$

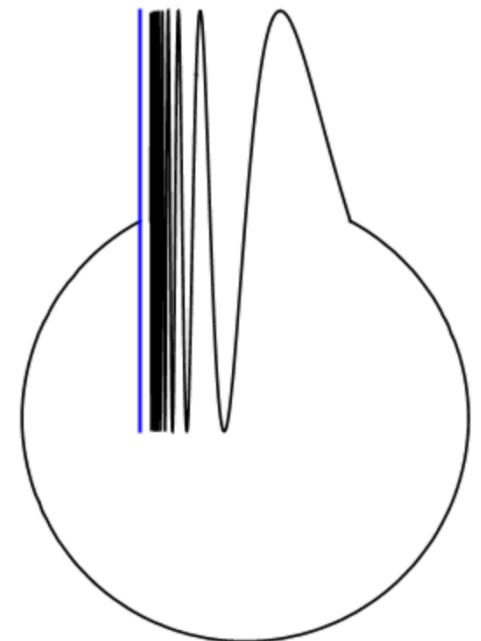
From Rips Complex

From Rips Complex

- ▶ Theorem [Homology Inference]
- ▶ Let $X \subset \mathbb{R}^d$ be a manifold and let $P \subset X$ be such that $d_H(X, P) \leq \epsilon$. If $2\epsilon \leq \alpha \leq \frac{1}{4}\sqrt{\frac{3}{5}}\rho(X)$, we have $\text{rank}(H_*(R^\alpha) \rightarrow H_*(R^{2\alpha})) = \dim H_*(X)$, in other words, the persistent Betti number of the VR filtration recovers the Betti number of X

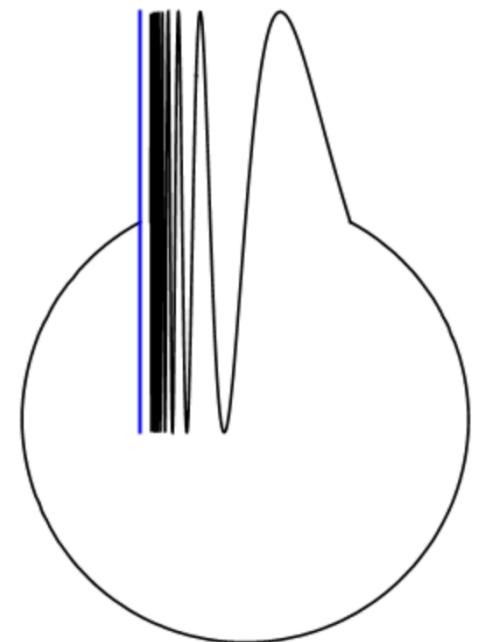
The Case of Compact Set

- ▶ $H_*(X^\lambda)$ may not be isomorphic to $H_*(X)$ even for sufficiently small $\lambda > 0$
- ▶ We no longer have isomorphism between $H_*(X)$ and $H_*(P^\alpha)$



The Case of Compact Set

- ▶ $H_*(X^\lambda)$ may not be isomorphic to $H_*(X)$ even for sufficiently small $\lambda > 0$
- ▶ We no longer have isomorphism between $H_*(X)$ and $H_*(P^\alpha)$
- ▶ We then only try to infer $H_*(X^\lambda)$ for small $\lambda > 0$ using pairs of spaces



The Case of Compact Set

Corollary A

Let $P \subset X$ be such that $d_H(X, P) \leq \epsilon$. If $2\epsilon \leq \alpha \leq \sqrt{\frac{3}{5}}\rho(X)$, we have $H_*(X) \cong H_*(P^\alpha)$

The Case of Compact Set

► In contrast to Corollary A, now we have the following (using critical point theory and Lemma B).

► Lemma C [*Chazal and Oudot 2008*]:

Let $P \subset \mathbb{R}^d$ be a finite set such that $d_H(X, P) < \epsilon$ for some $\epsilon < \frac{1}{4}wfs(X)$. Then for all $\alpha, \beta \in [\epsilon, wfs(X) - \epsilon]$ such that $\beta - \alpha \geq 2\epsilon$, and for all $\lambda \in (0, wfs(X))$, we have $H(X^\lambda) \cong \text{image}(i_*)$, where $i_*: H(P^\alpha) \rightarrow H(P^\beta)$ is the homomorphism between homology groups induced by the canonical inclusion $i: P^\alpha \rightarrow P^\beta$.

The Case of Compact Set

- ▶ Lemma C [*Chazal and Oudot 2008*]:

Let $P \subset \mathbb{R}^d$ be a finite set such that $d_H(X, P) < \epsilon$ for some $\epsilon < \frac{1}{4}wfs(X)$. Then for all $\alpha, \beta \in [\epsilon, wfs(X) - \epsilon]$ such that $\beta - \alpha \geq 2\epsilon$, and for all $\lambda \in (0, wfs(X))$, we have $H(X^\lambda) \cong \text{image}(i_*)$, where $i_*: H(P^\alpha) \rightarrow H(P^\beta)$ is the homomorphism between homology groups induced by the canonical inclusion $i: P^\alpha \rightarrow P^\beta$.

- ▶ $X^\lambda \simeq X^{\lambda'}$ for $\lambda, \lambda' \in (0, wfs(X))$
- ▶ $X^{\alpha-\epsilon} \subseteq P^\alpha \subseteq X^{\alpha+\epsilon} \subseteq P^\beta \subseteq X^{\beta+\epsilon}$
- ▶ Using Nerve theorem, this can be transformed into a result regarding Čech filtrations

The Case of Compact Set

The Case of Compact Set

- ▶ Theorem [Homology Inference] [*Chazal and Oudot 2008*]:

Let $P \subset \mathbb{R}^d$ be a finite set such that $d_H(X, P) < \epsilon$ for some $\epsilon < \frac{1}{9}wfs(X)$. Then for all

$\alpha \in \left[2\epsilon, \frac{1}{4}(wfs(X) - \epsilon)\right]$ all $\lambda \in (0, wfs(X))$, we have $H(X^\lambda) \cong \text{image}(j_*)$,

where j_* is the homomorphism between homology groups induced by canonical inclusion $j: R^\alpha \rightarrow R^{4\alpha}$.

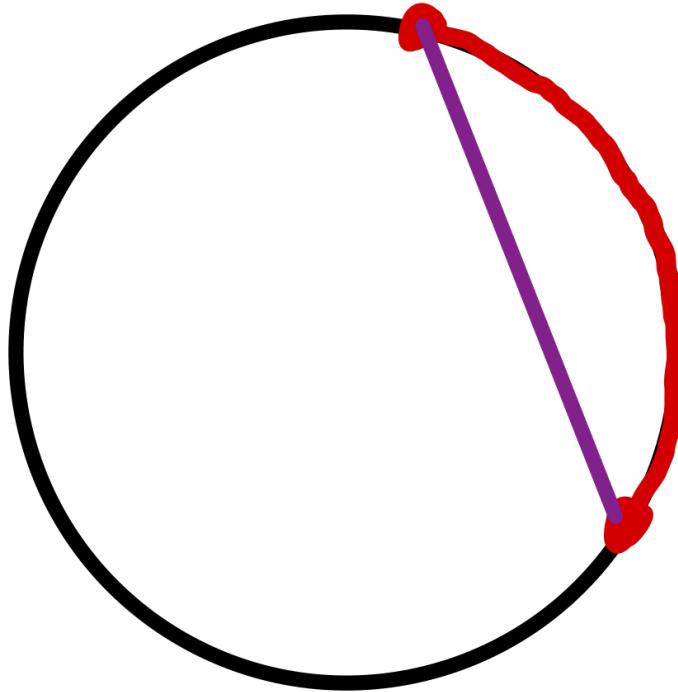
Summary of Homology Inference

- ▶ NSW theorem or the critical point theory of d_X gives us information on topological change in the filtration $\{X^r\}$
- ▶ The Hausdorff distance gives rise to interleaving between $\{X^r\}$ and $\{P^r\}$
- ▶ Nerve Theorem relates Čech filtration $\{C^r\}$ with $\{P^r\}$ through isomorphism
- ▶ Rips and Čech are related through interleaving
- ▶ Both are related to $\{X^r\}$ through interleaving
- ▶ Homology of X can be then inferred from Rips or Čech filtrations of P

Homology inference for manifolds

Hausmann's theorem

- ▶ Let M be a Riemannian manifold (i.e., a manifold with a metric structure)



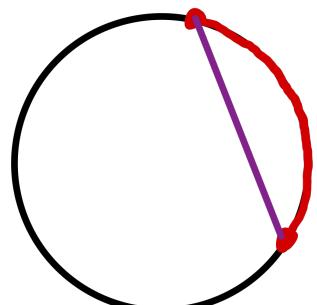
Hausmann's theorem

- ▶ Let M be a Riemannian manifold (i.e., a manifold with a metric structure)

- ▶ Consider

$$Rips^r(M) = \left\{ (p_0, \dots, p_k) \mid d(p_i, p_j) < r, \forall i, j \text{ and } \forall p_0, \dots, p_k \in M \right\}$$

- ▶ For r sufficiently small, then $Rips^r(M)$ is homotopy equivalent to M



Latschev's theorem - sampling version

- ▶ Let M be a Riemannian manifold (i.e., a manifold with a metric structure)
- ▶ For any metric space X such that $d_{GH}(X, M) \leq \epsilon$ sufficiently small,
- ▶ Then $Rips^\epsilon(X)$ is homotopy equivalent to M

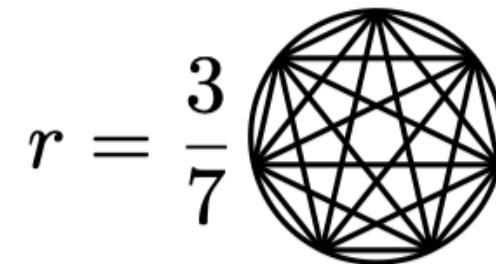
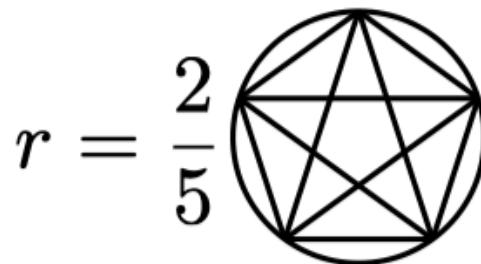
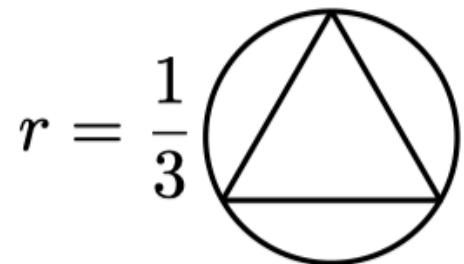
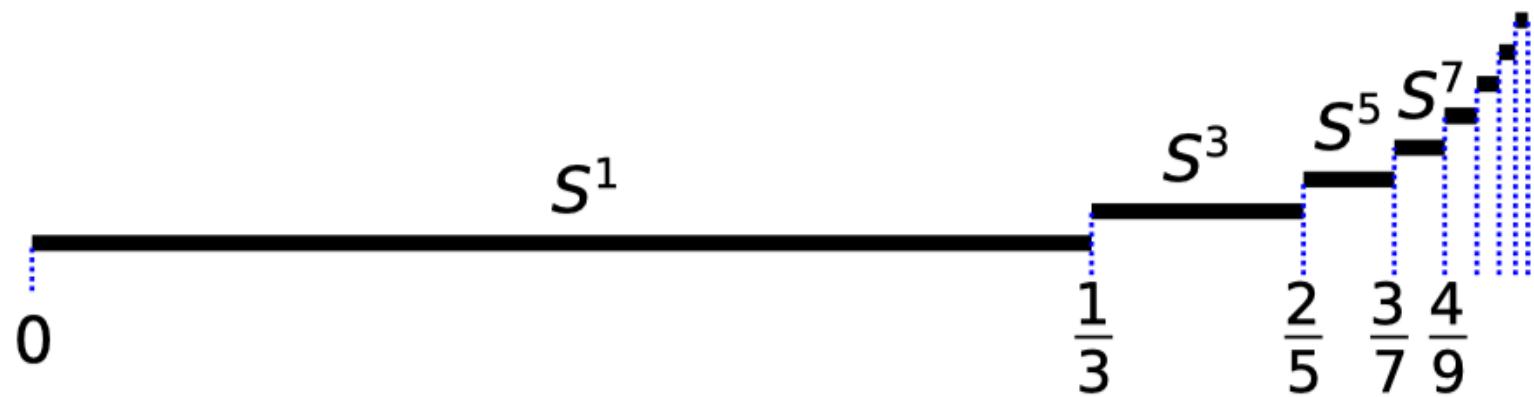
Latschev's theorem - sampling version

- ▶ Let M be a Riemannian manifold (i.e., a manifold with a metric structure)
- ▶ For any metric space X such that $d_{GH}(X, M) \leq \epsilon$ sufficiently small,
- ▶ Then $Rips^\epsilon(X)$ is homotopy equivalent to M

Can we say anything “persistent”?

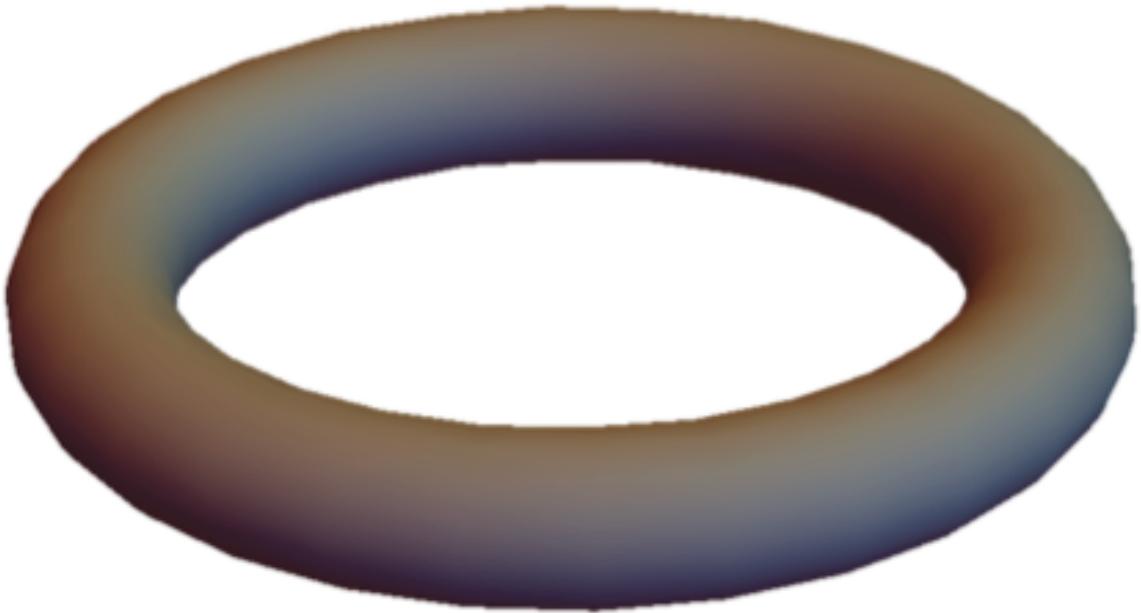
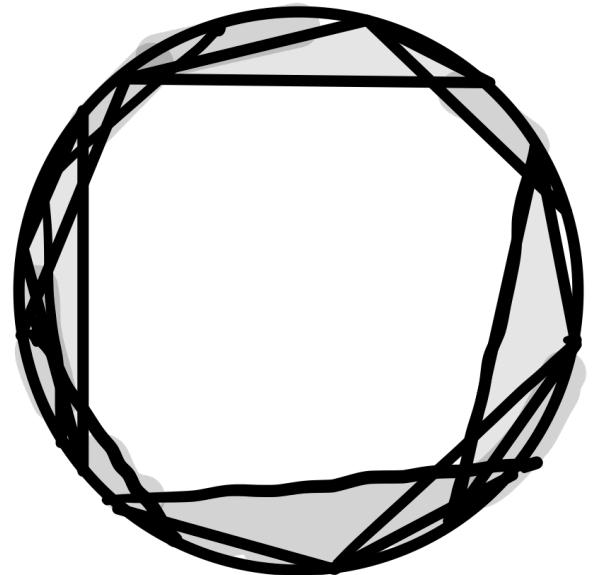
VR complex of \mathbb{S}^1

- ▶ $Rips^r(\mathbb{S}^1) \simeq \mathbb{S}^{2l+1}, \frac{l}{2l+1} < r \leq \frac{l+1}{2l+3}$

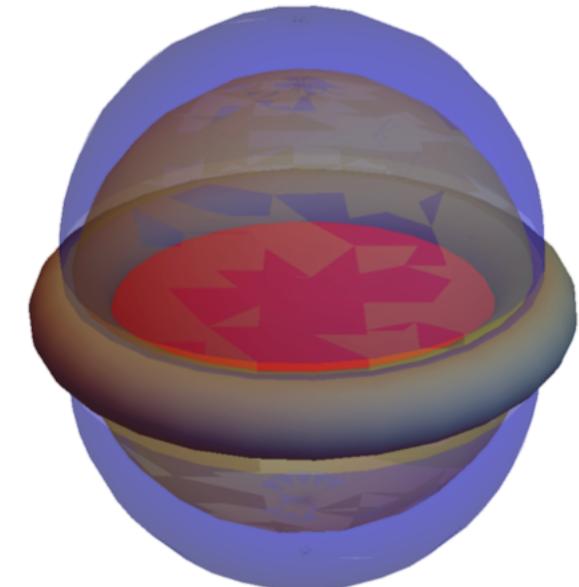
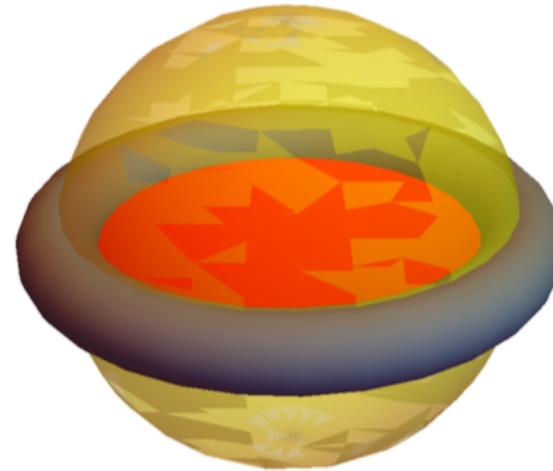
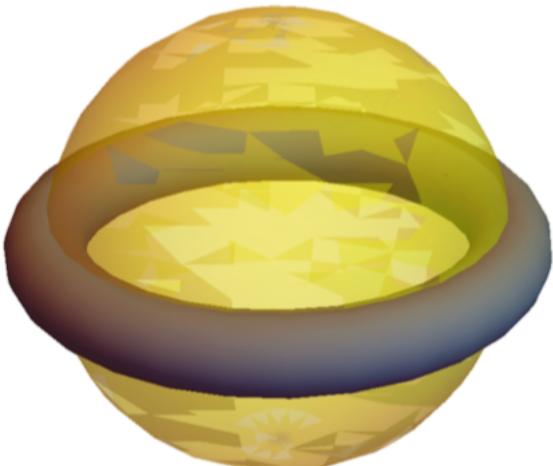
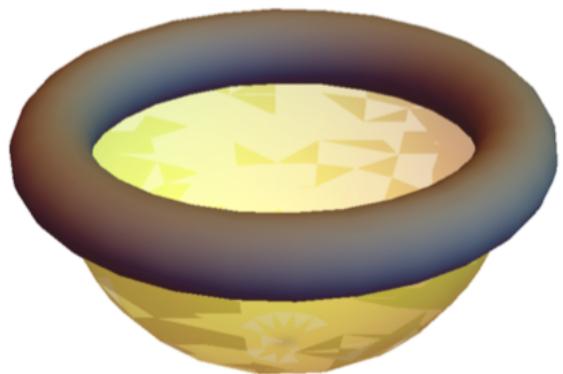
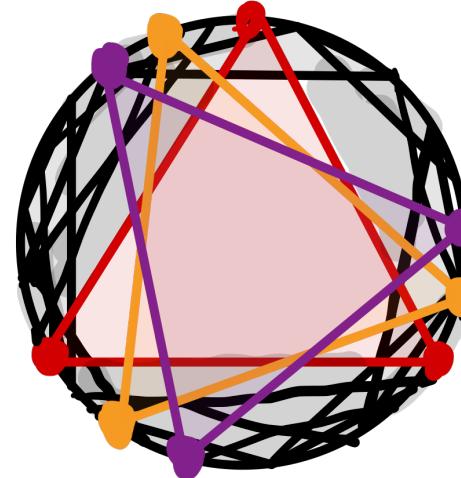
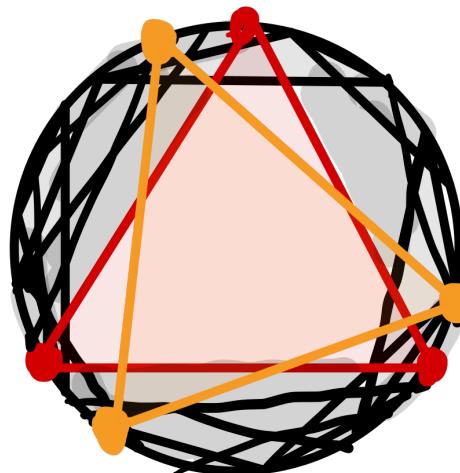
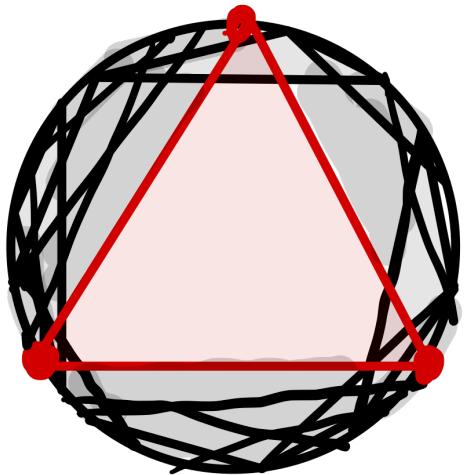


Courtesy of Henry Adams

$$r \leq 1/3$$

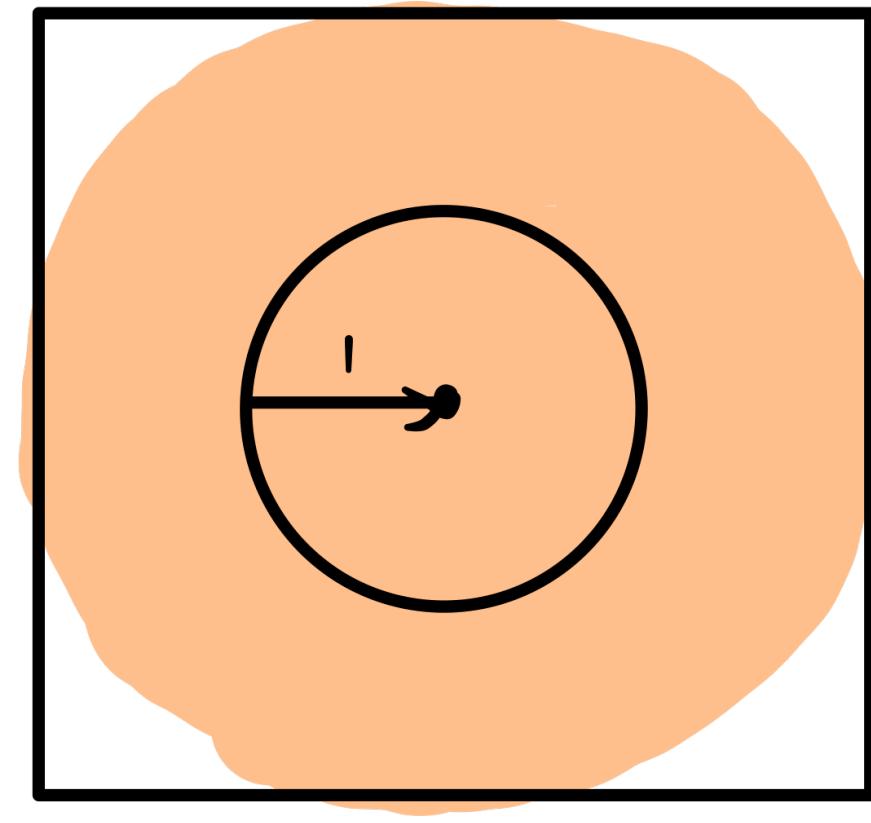
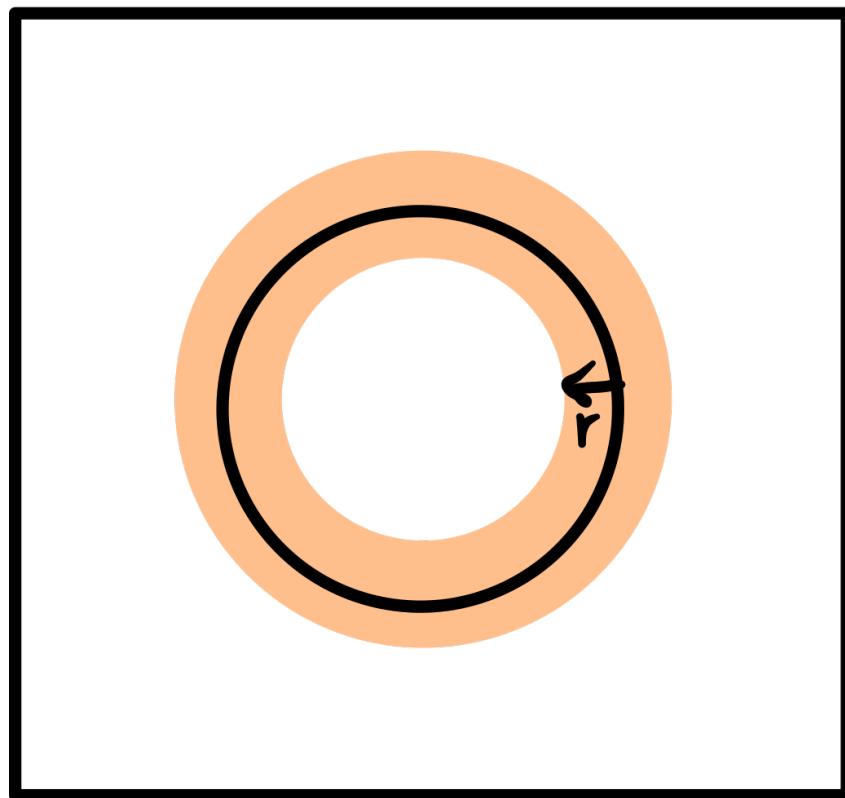


$r > 1/3$



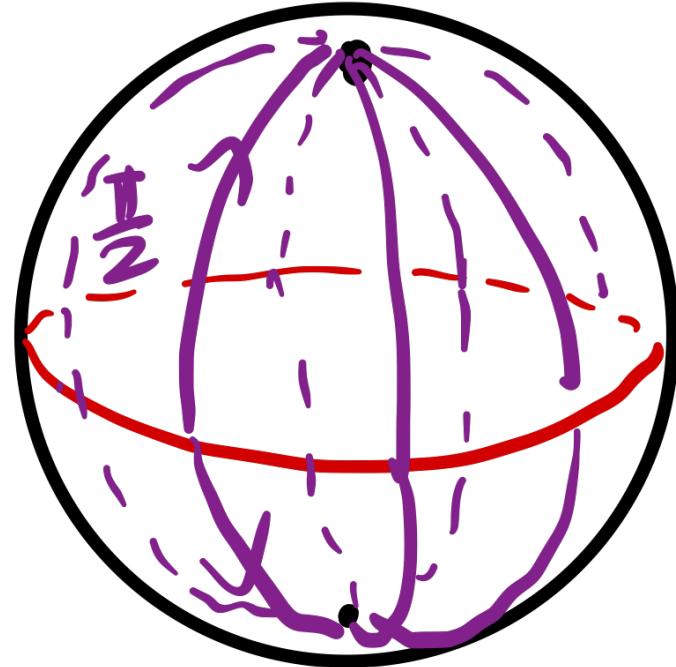
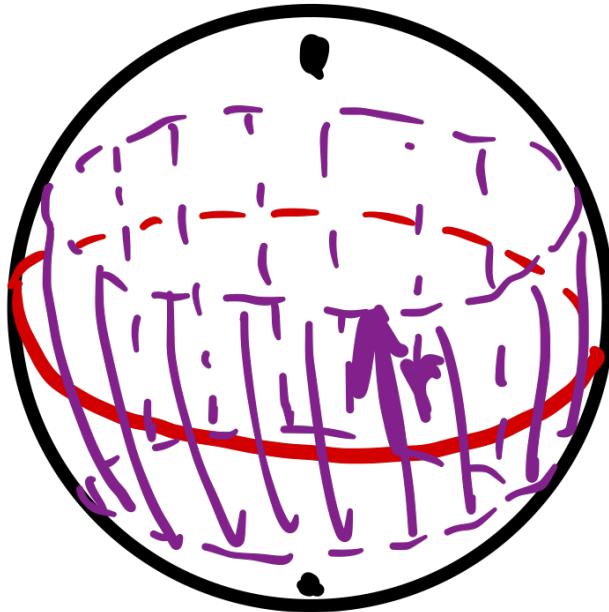
Courtesy of Henry Adams

- ▶ Can we interpret this using certain offset idea?



Filling radius

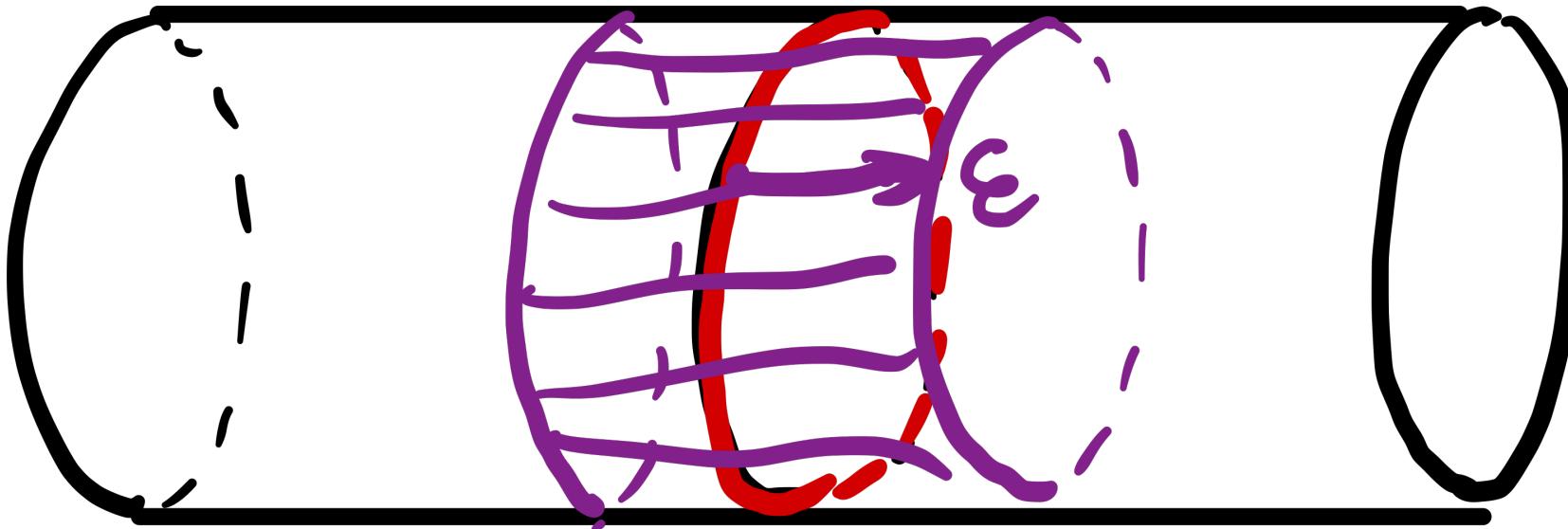
- ▶ $\text{FillRad}(M, X) = \inf\{\epsilon : M^\epsilon \subset X \text{ has trivial top dim homology}\}$



$$\text{FillRad}(\mathbb{S}^1, \mathbb{S}^2) = \pi/2$$

Filling radius

- ▶ $\text{FillRad}(M, X) = \inf\{\epsilon : M^\epsilon \subset X \text{ has trivial top dim homology}\}$



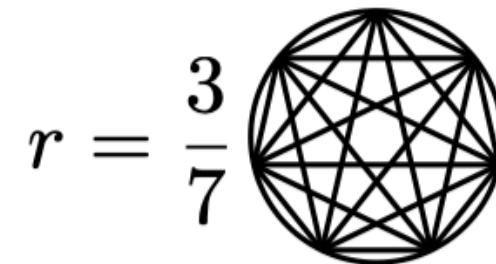
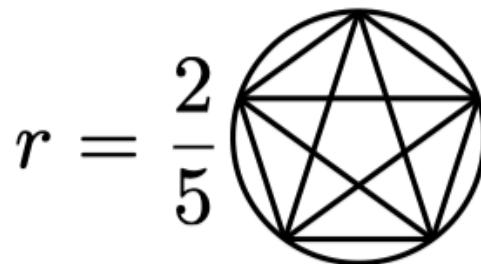
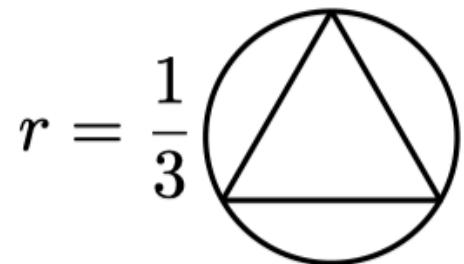
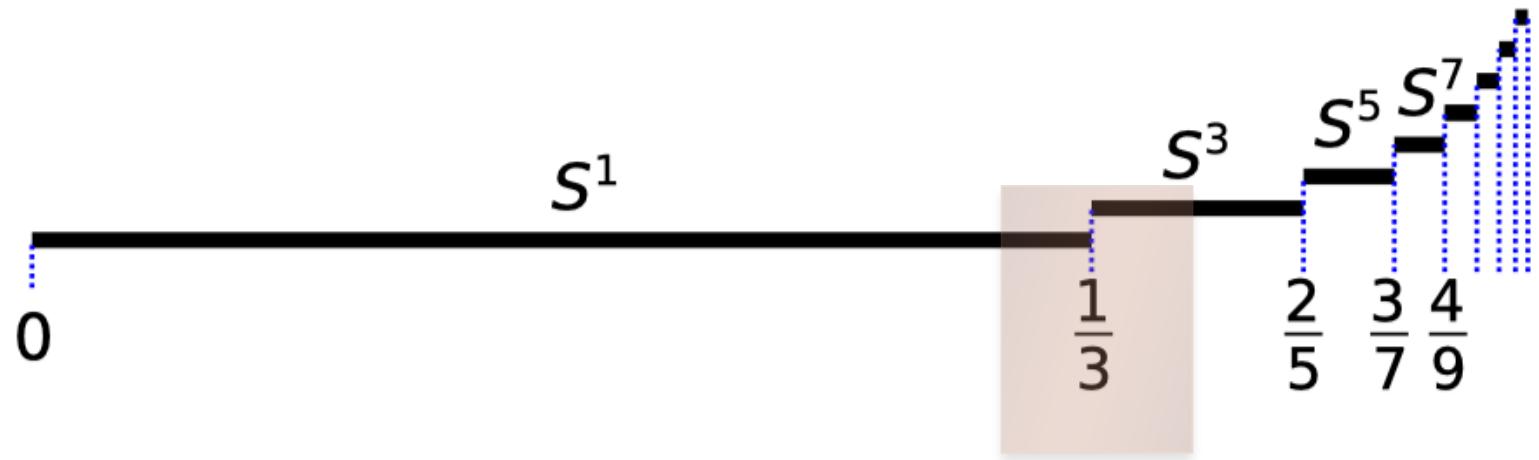
$$\text{FillRad}(\mathbb{S}^1, X) = \infty$$

Filling radius

- ▶ $\text{FillRad}(M, X) = \inf\{\epsilon : M^\epsilon \subset X \text{ has trivial top dim homology}\}$
- ▶ Filling radius of M : $\text{FillRad}(M) := \inf_X \text{FillRad}(M, X)$
- ▶ $\text{FillRad}(\mathbb{S}^1) = \pi/3$ and $\text{FillRad}(\mathbb{S}_1^1) = 1/6$

VR complex of \mathbb{S}^1

- ▶ $Rips^r(\mathbb{S}^1) \simeq \mathbb{S}^{2l+1}, \frac{l}{2l+1} < r \leq \frac{l+1}{2l+3}$

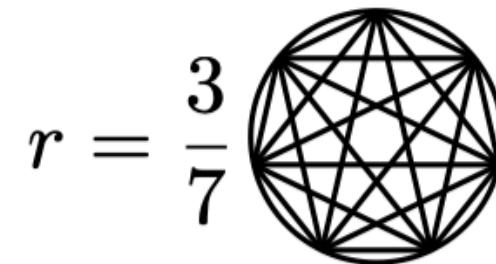
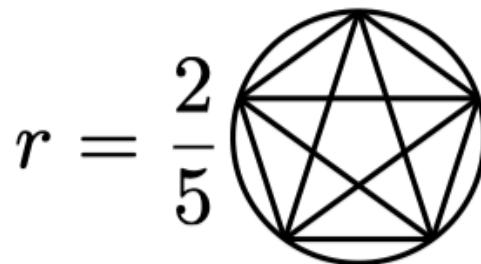
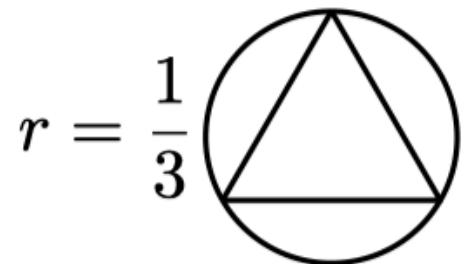
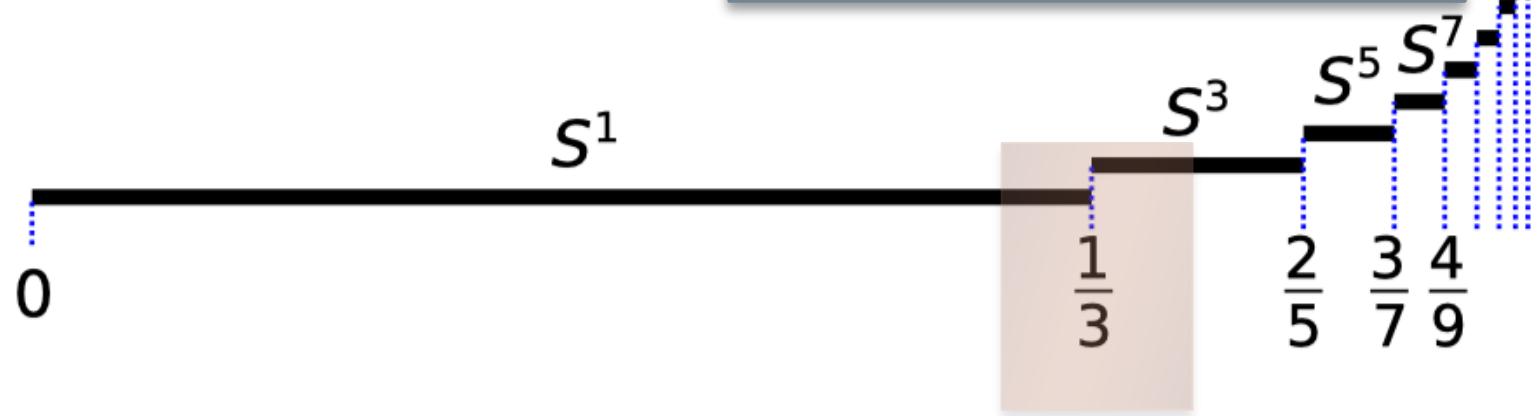


Courtesy of Henry Adams

VR complex of \mathbb{S}^1

► $Rips^r(\mathbb{S}^1) \simeq \mathbb{S}^{2l+1}, \frac{l}{2l+1} < r \leq \frac{l+1}{2l+1}$

$$1/3 = 2 * FillRad(\mathbb{S}_1^1)$$



Courtesy of Henry Adams

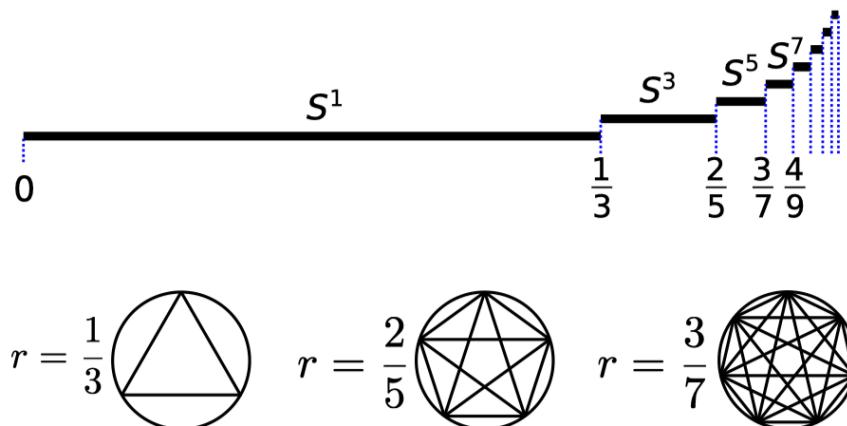
Filling radius and VR filtration

- ▶ [Lim et al. 2022]

Proposition 9.4. *Let M be a closed connected n -dimensional Riemannian manifold. Then,*

$$(0, 2 \text{FillRad}(M)] \in \text{barc}_n^{\text{VR}}(M; \mathbb{F}),$$

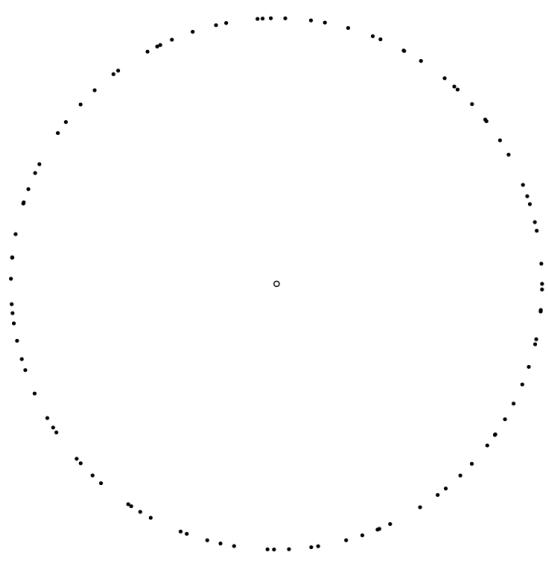
where \mathbb{F} is an arbitrary field if M is orientable, and $\mathbb{F} = \mathbb{Z}_2$ if M is non-orientable. Moreover, this is the unique interval in $\text{barc}_n^{\text{VR}}(M; \mathbb{F})$ starting at 0.



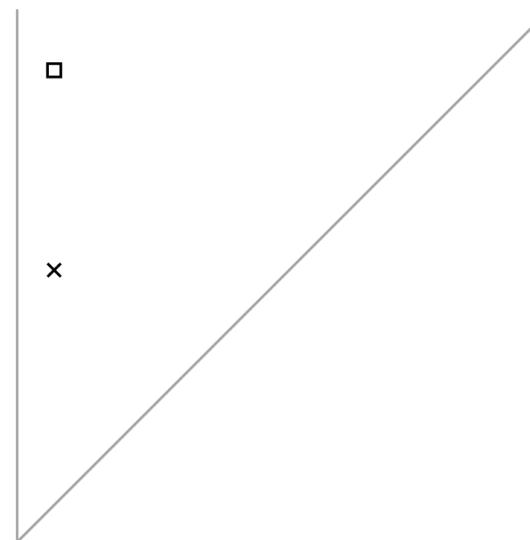
Section 2: Handling of Noise

Noise

- ▶ Previous approach can 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



(a) Samples



(b) Diagrams

Courtesy of Bendich et al.

Noise

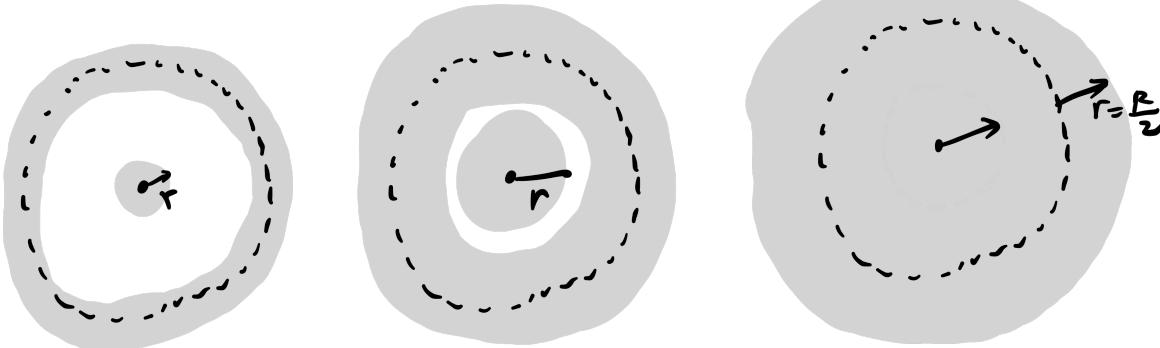
- ▶ Previous approach can handle Hausdorff type noise
 - ▶ Where noise is within a tubular neighborhood of X
- ▶ How about more general noise?
 - ▶ E.g, Gaussian noise, background noise
- ▶ Distance to measure framework
 - ▶ *[Chazal, Cohen-Steiner, Mérigot, 2011]*

Main idea

$$d_H \left(\left(\text{solid circle} \right), \left(\text{dashed circle} \right) \right) \approx R$$

$$d_W \left(\left(\text{solid circle} \right), \left(\text{dashed circle} \right) \right) \approx O\left(\frac{1}{n}\right) \ll 1$$

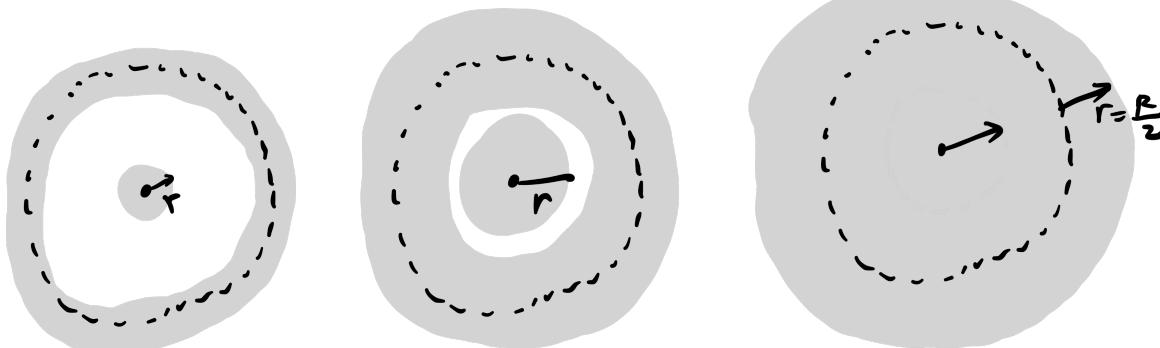
The sub level set of the distance to set d_X



Main idea

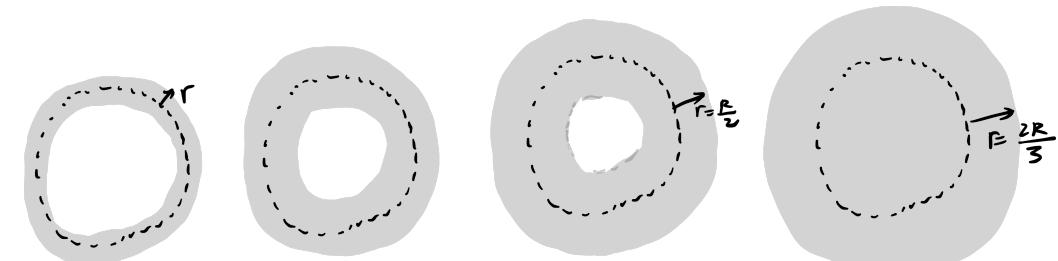
$$d_H \left(\left(\text{solid circle} \right), \left(\text{dashed circle} \right) \right) \approx R$$

The sub level set of the distance to set d_X



$$d_W \left(\left(\text{solid circle} \right), \left(\text{dashed circle} \right) \right) \approx O\left(\frac{1}{n}\right) \ll 1$$

The sub level set of the “distance to measure”



Main Idea

- ▶ The work of [*Chazal, Cohen-Steiner, Mérigot, 2011*]
- ▶ A new distance function $d_{\mu,m}$ to a probability measure μ (ie., distance to measures) to replace the role of distance function d_X .
- ▶ Show that the two distance functions are close (in L_∞ norm)
- ▶ Topological inference follows from some stability results or the interleaving sequences

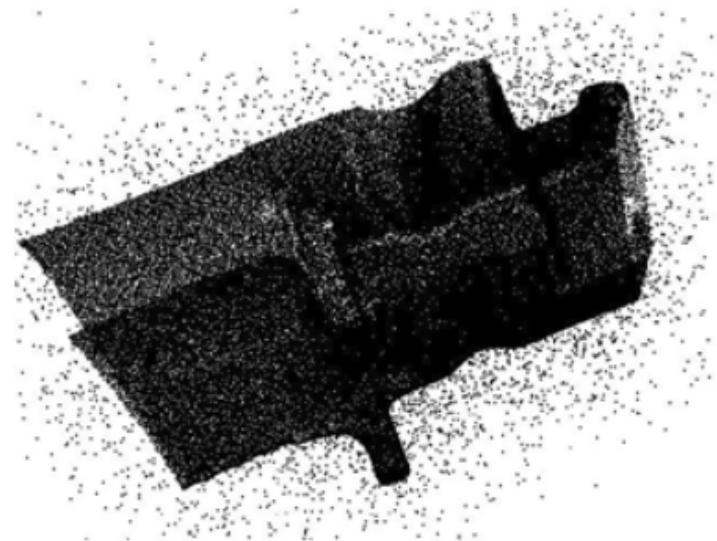
Overview

- ▶ **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



Courtesy of Chazal et al 2011

Definitions

- ▶ μ : a probability measure on \mathbb{R}^d ; $\mu(\mathbb{R}^d) = 1$
- ▶ $0 < m, m_0 < 1$: **mass** parameters
- ▶ $\delta_{\mu,m}(x) := \inf\{r > 0; \mu(\bar{B}(x, r)) > m\}$

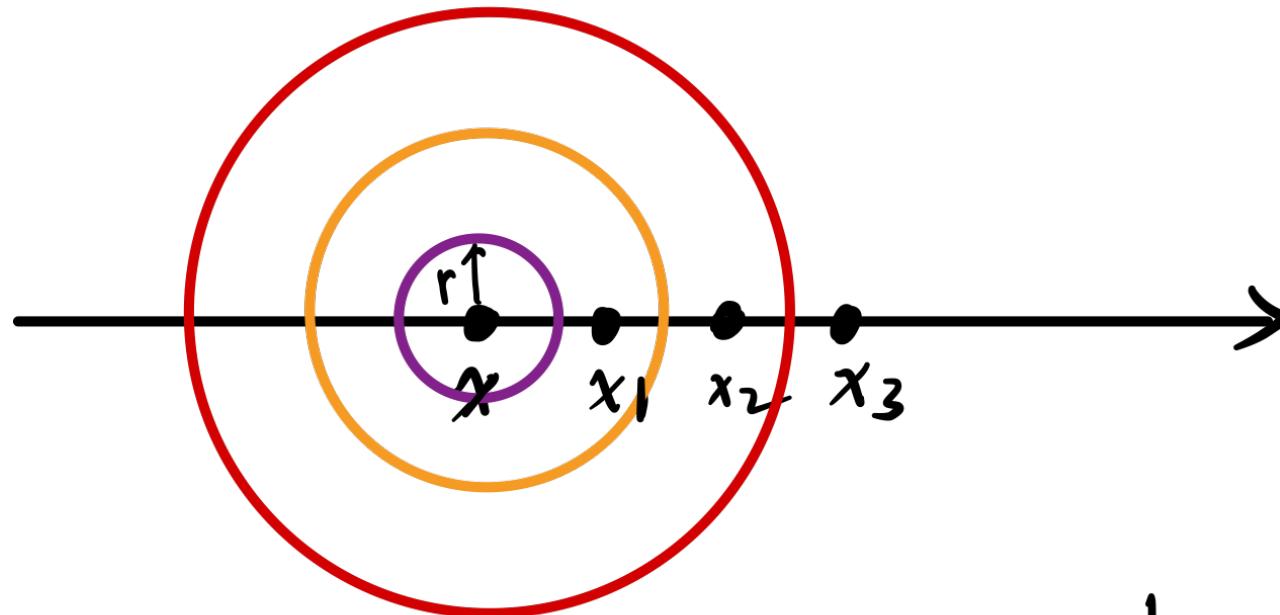
where $\bar{B}(x, r)$ is the closed Euclidean ball at x
- ▶ That is, $\delta_{\mu,m}(x)$ is the radius of the ball necessary in order to enclose mass m

Discrete setting

- ▶ Given a set of points $P \subset \mathbb{R}^d$, consider the empirical probability measure
$$\mu_P = \frac{1}{n} \sum_{p \in P} \delta_p$$
 where intuitively every point has measure $\frac{1}{n}$
- ▶ For any integer k , we let $kNN(x)$ denote the set of k nearest neighbors of x in P
 - ▶ Denote $kNN(x) = \{x_1, \dots, x_k\}$ such that $\|x - x_i\| \leq \|x - x_{i+1}\|$
 - ▶ In this case if $\frac{k-1}{n} \leq m < \frac{k}{n}$, then
 - ▶ $\delta_{\mu_P, m} = \|x - x_k\|$

Discrete setting

- ▶ $\delta_{\mu,m}(x) := \inf\{r > 0; \mu(\bar{B}(x, r)) > m\}$



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

Definitions

- ▶ μ : a probability measure on \mathbb{R}^d ; $\mu(\mathbb{R}^d) = 1$
- ▶ $0 < m, m_0 < 1$: *mass* parameters
- ▶ $\delta_{\mu,m}(x) := \inf\{r > 0; \mu(\bar{B}(x, r)) > m\}$

where $\bar{B}(x, r)$ is the closed Euclidean ball at x

- ▶ That is, $\delta_{\mu,m}(x)$ is the radius of the ball necessary in order to enclose mass m
- ▶ *Distance to measure* d_{μ,m_0} :

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

Distance to Measures

- ▶ *Distance to measure* d_{μ, m_0} :

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

- ▶ Intuitive, $d_{\mu, m}(x)$ averages distance within a range and is more robust to noise.
- ▶ Note this distance depends on a mass parameter m_0

Discrete setting

- ▶ Given a set of points $P \subset \mathbb{R}^d$, consider the empirical probability measure $\mu_P = \frac{1}{n} \sum_{p \in P} \delta_p$ where intuitively every point has measure $\frac{1}{n}$
- ▶ In this case, if $m = k/n$, then
 - ▶ $d_{\mu_P, m}(x) = \sqrt{\frac{1}{k} \sum_{q \in kNN(x)} \|x - q\|^2} =: d_{\mu_P, k}(x)$

Wasserstein Distance

- ▶ A *transport plan* between two probability measures μ, ν on \mathbb{R}^d is a probability measure π on $\mathbb{R}^d \times \mathbb{R}^d$ such that for every $A, B \subseteq \mathbb{R}^d$, $\pi(A \times \mathbb{R}^d) = \mu(A)$ and $\pi(\mathbb{R}^d \times B) = \nu(B)$.
- ▶ The *p-cost* of a transport plan π is:

$$C_p(\pi) = \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \pi(dx \times dy) \right)^{1/p}$$

- ▶ The *Wasserstein distance* of order p between μ, ν on \mathbb{R}^d with finite p -moment
 - ▶ $W_p(\mu, \nu) =$ the minimum p-cost $C_p(\pi)$ of any transport plan π between μ and ν .

Wasserstein Distance

- ▶ The Wasserstein distance can be generalized to compare two measures μ and ν with the same total mass (not necessarily $=1$)
- ▶ A measure ν is a submeasure of μ if $\nu(B) \leq \mu(B)$ for any measurable set B . Let $Sub_m(\mu)$ denote the set of all submeasures of μ with total mass m

- ▶
$$d_{\mu,m_0}(x) = \min_{\nu \in Sub_{m_0}(\mu)} \frac{1}{\sqrt{m_0}} W_2(m_0 \delta_x, \nu)$$

Properties

Properties

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

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

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

Properties

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

Let μ, μ' 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 μ, μ' 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

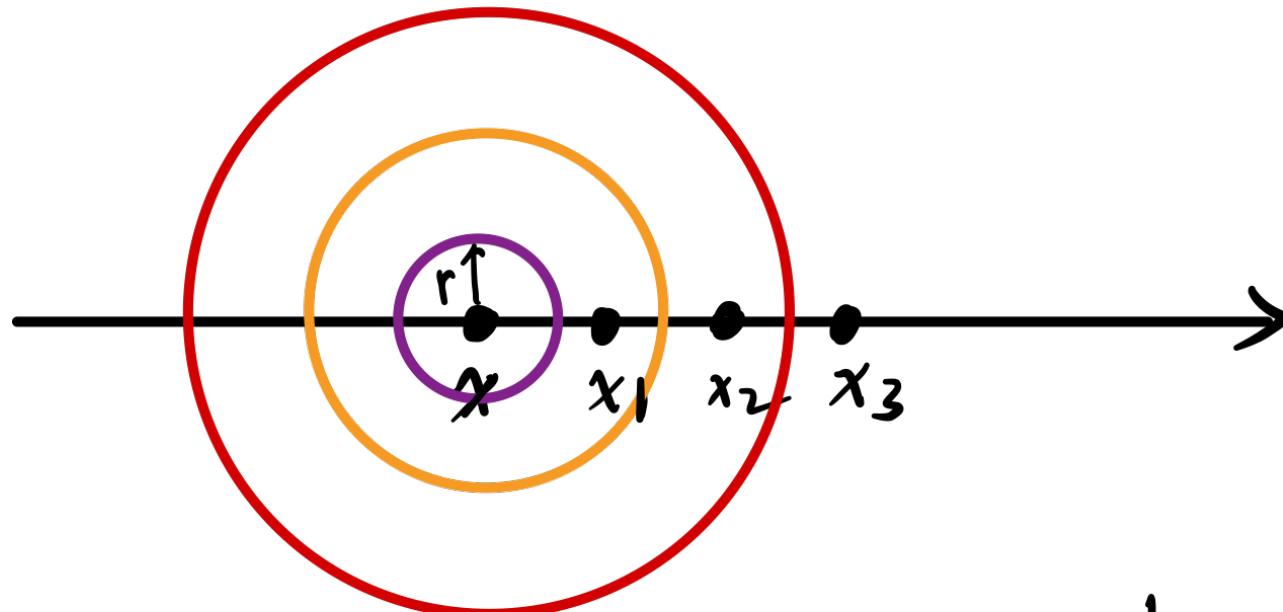
- Theorem [Approximation Distance]:

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

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

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

Relation to Distance Function



- $d_{\mu, m_0}^2(x) = \|x - x_1\|^2 = d_X(x)$
when $m_0 \in (0, 1/3)$

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

Relation to Distance Function

- ▶ Now suppose P is sampled from, not compact X , but a probabilistic measure μ on \mathbb{R}^d concentrated on X .
 - ▶ Consider P as a noisy sample of X
- ▶ Let ν_X denote the uniform measure on X

- ▶ Theorem [Approximation Distance]:

$$\left\| d_X - d_{\mu, m_0} \right\|_\infty \leq C(X)^{-\frac{1}{k}} m_0^{\frac{1}{k}} + \frac{1}{\sqrt{m_0}} W_2(\mu, \nu_X)$$

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

Topological inference

Theorem [Niyogi, Smale, Weinberger]

Let $P \subset X$ be such that $d_H(X, P) \leq \epsilon$. If $2\epsilon \leq \alpha \leq \sqrt{\frac{3}{5}}\rho(X)$,
there is a deformation retraction from P^α to 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 $\text{reach}_\alpha(d_K) \geq R$ 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*

$$W_2(\mu, \mu') \leq \frac{R\sqrt{m_0}}{5 + 4/\alpha^2} - C(\mu)^{-1/k} m_0^{1/k+1/2}.$$

Noisy sample

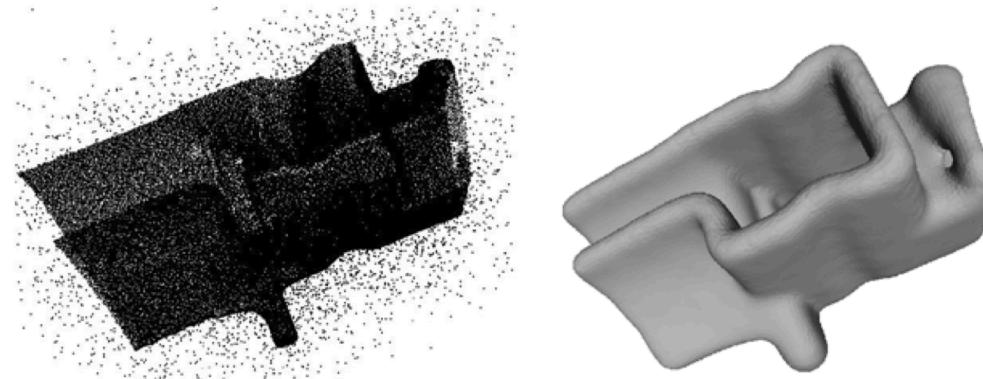
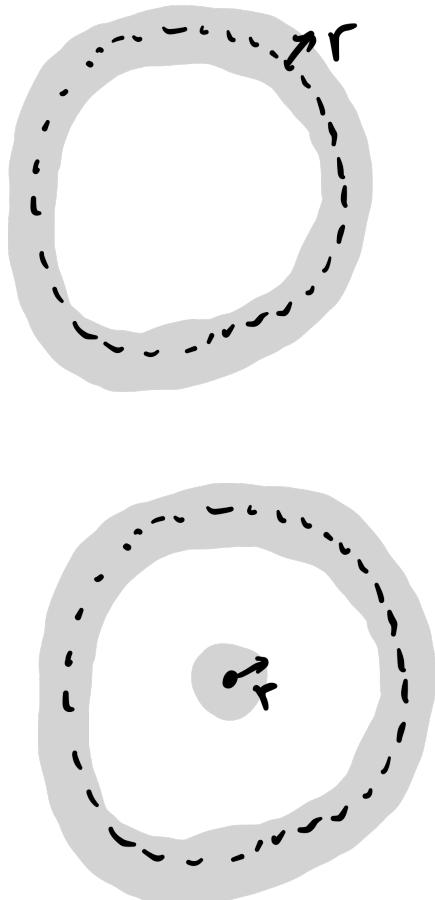
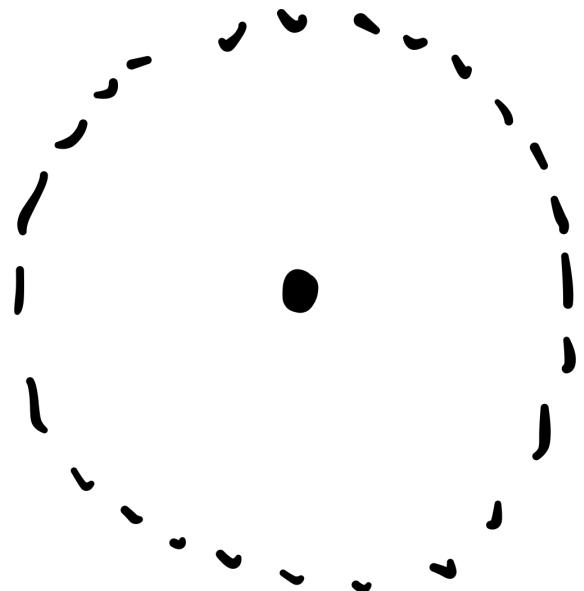


Fig. 1 *Left:* a point cloud sampled on a mechanical part to which 10% of outliers (uniformly sampled in a box enclosing the model) have been added. *Right:* the reconstruction of an isosurface of the distance function d_{μ_C, m_0} to the uniform probability measure on this point cloud

Sublevel set of distance to measure

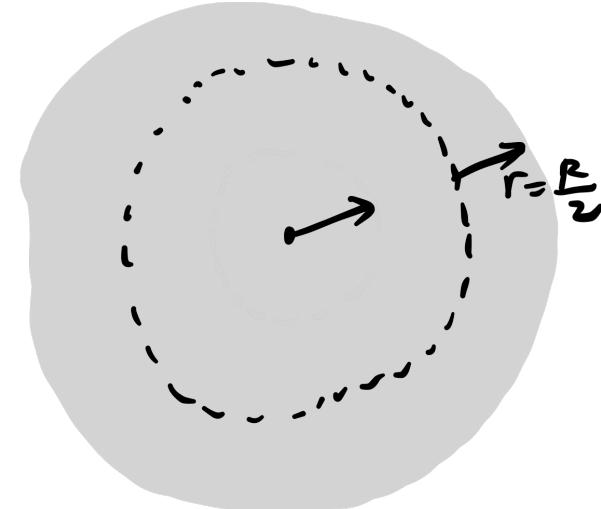
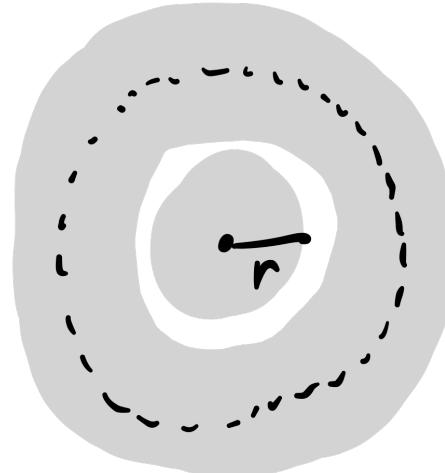
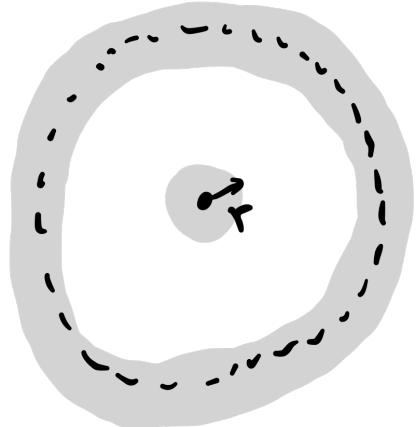


$$k=3 \\ d_{\mu,k}^{(x)} = \sqrt{\frac{1}{k} \sum_{q \in N(x)} d^2(x,q)} \leq r$$

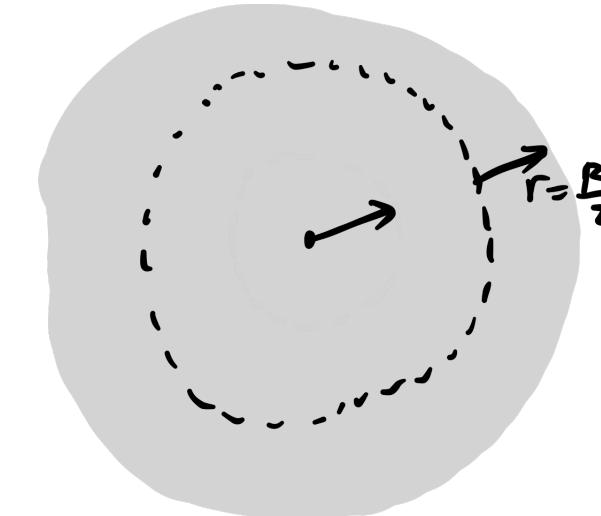
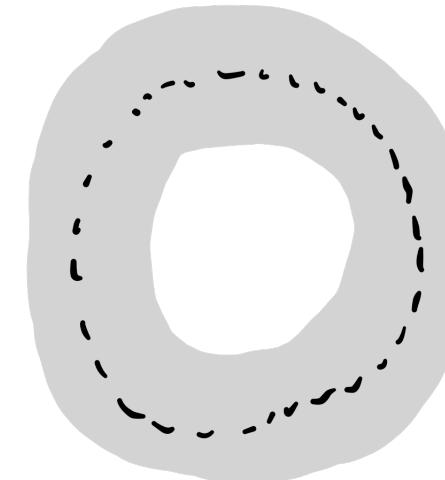
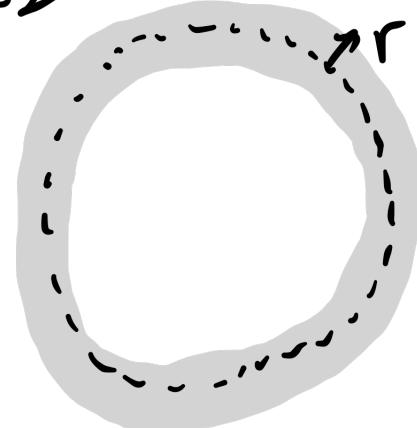
$$d_x^{-1}(-\infty, r])$$

Sublevel set of distance to measure

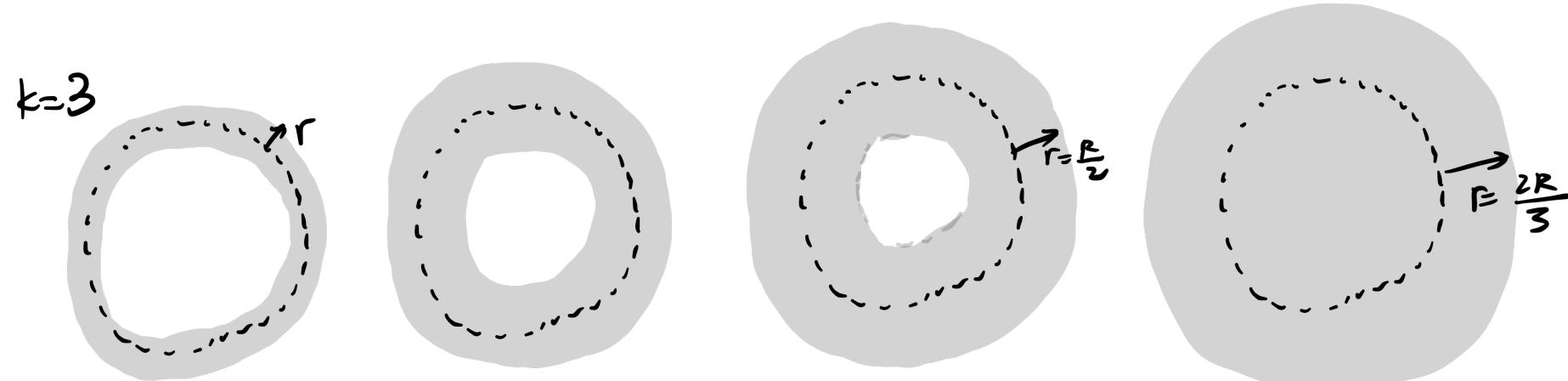
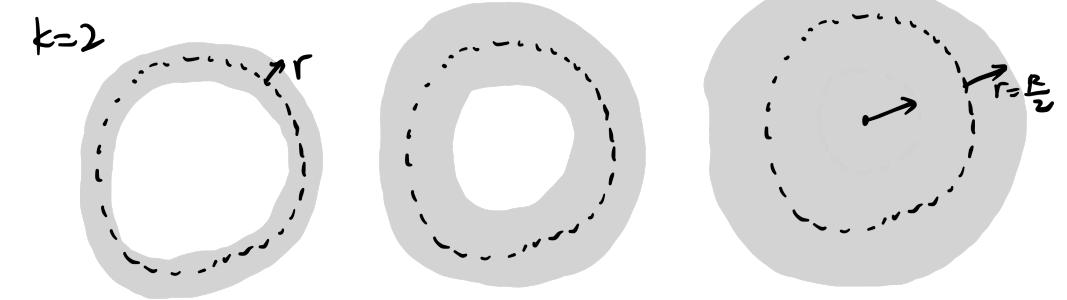
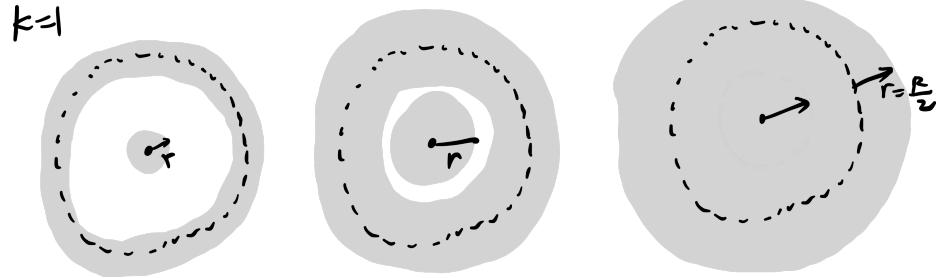
$k=1$



$k=2$



Sublevel set of distance to measure



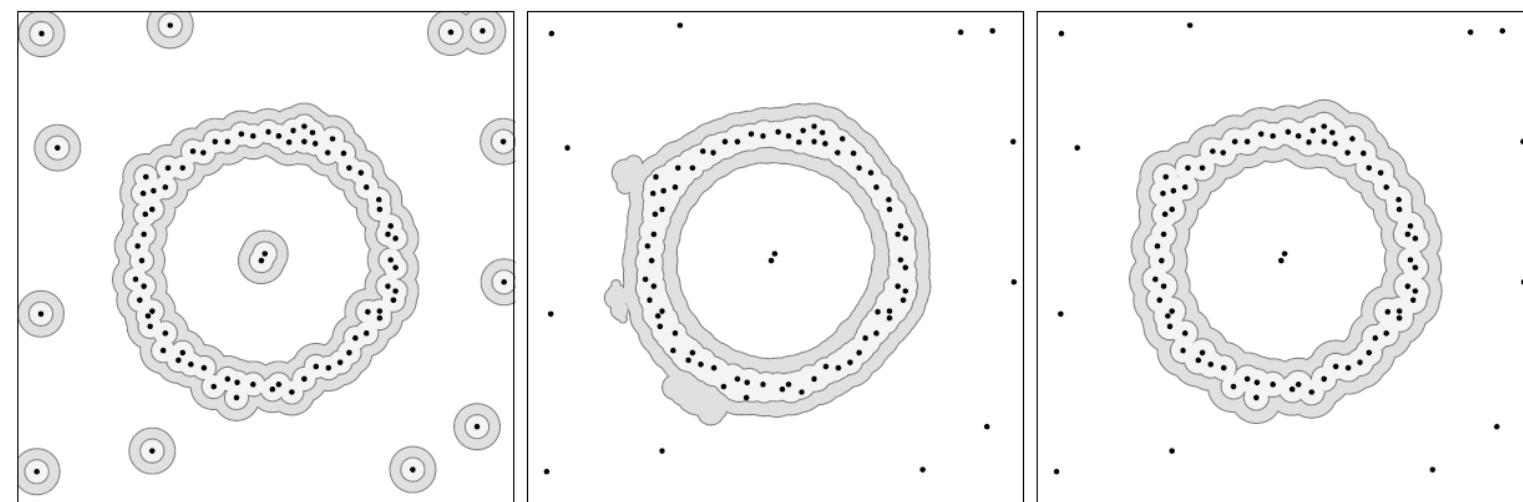
Approximate distance to measure using weighted Rips

How to use the distance to measures?

- ▶ Previous theorem suggests that using distance-to-measure, instead of distance can be used to approximate topology.
- ▶ In practice, using union of weighted balls of non-uniform radius to approximate hidden space, so that bad points have lower contribution!
- ▶ One way to achieve this is through the power-distance
 - ▶ [Buchet et al., 2016]

How to use the distance to measures?

- ▶ Previous theorem suggests that using distance-to-measure, instead of distance can be used to approximate topology.
- ▶ In practice, using union of weighted balls of non-uniform radius to approximate hidden space, so that bad points have lower contribution!
- ▶ One way to achieve this is through the power-distance
 - ▶ [Buchet et al., 2016]



Definition 4.1. Given a metric space \mathbb{X} , a set P and a function $w : P \rightarrow \mathbb{R}$, 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 .

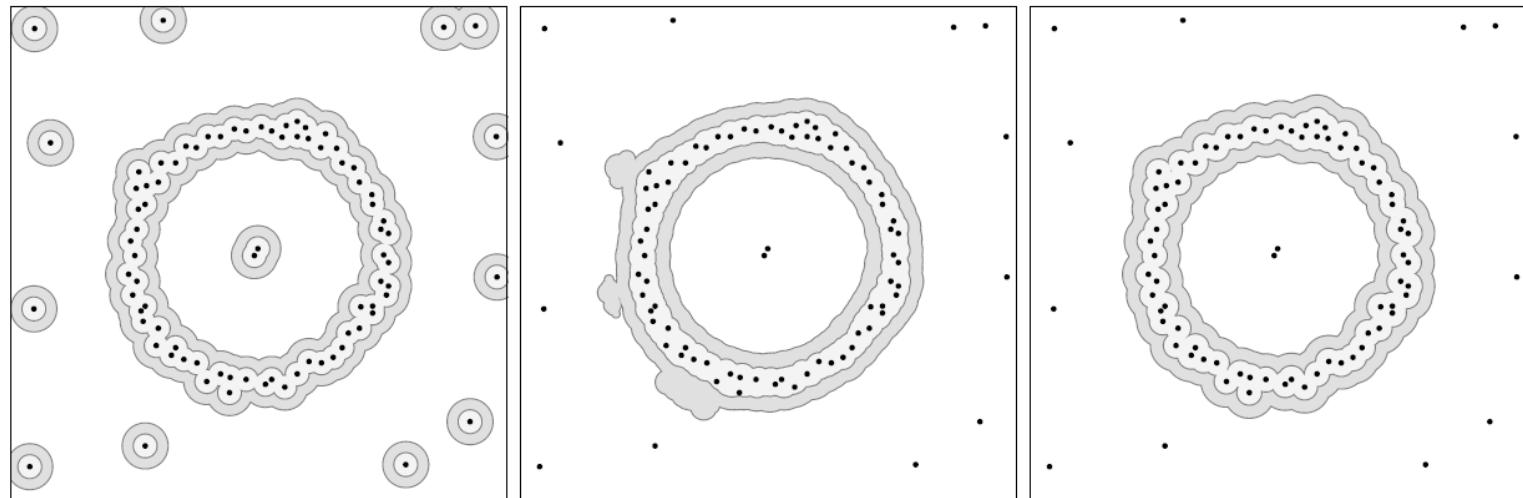
- ▶ Note when $w_p \equiv 0$, then $f(x) = d_P(x)$
- ▶ Set $w_p = d_{\mu,m}(p)$, the corresponding f is an approximation of $d_{\mu,m}$
- ▶ So the sub level set filtration of f is an approximation of the one for $d_{\mu,m}$
- ▶ Let $r_p(\alpha) = \sqrt{\alpha^2 - w_p^2}$. Then, $f^{-1}((-\infty, \alpha]) = \cup_{p \in P} B(p, r_p(\alpha))$
- ▶ This corresponds to a weighted Čech complex which can be approximated by a weighted Rips complex
- ▶ $wR^\alpha(P) = \{\sigma = (p_{i_0}, \dots, p_{i_s}) \mid d(p_{i_j}, p_{i_{j'}}) \leq r_{p_{i_j}}(\alpha) + r_{p_{i_{j'}}}(\alpha), \forall j \neq j' \in [0, s]\}$

Weighted Rips Filtration

- ▶ Weighted Rips filtration
 - ▶ $w\mathcal{F}: wR^{\alpha_0}(P) \hookrightarrow wR^{\alpha_1}(P) \hookrightarrow \dots \hookrightarrow wR^{\alpha_m}(P)$
- ▶ Then compute the persistence diagram induced by $w\mathcal{F}$, which is an approximation of $Dgm(d_{\mu,m})$

Weighted Rips Filtration

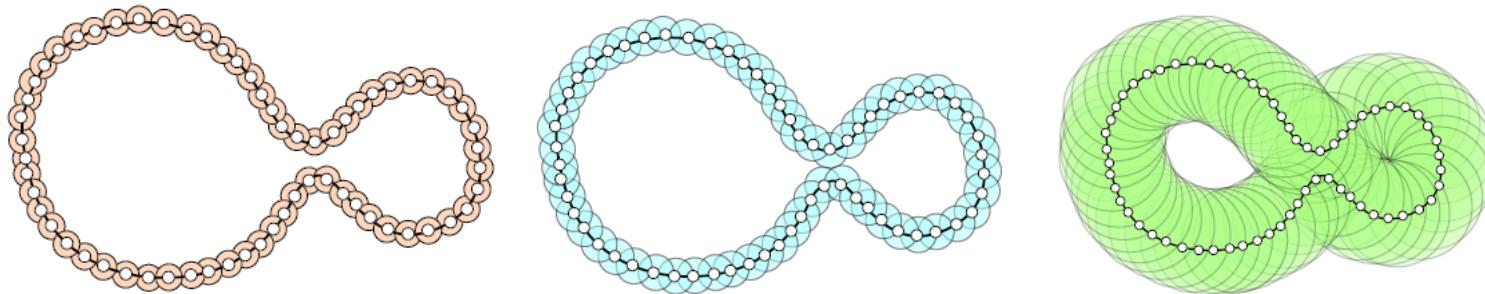
- ▶ Weighted Rips filtration
 - ▶ $w\mathcal{F}: wR^{\alpha_0}(P) \hookrightarrow wR^{\alpha_1}(P) \hookrightarrow \dots \hookrightarrow wR^{\alpha_m}(P)$
- ▶ Then compute the persistence diagram induced by $w\mathcal{F}$, which is an approximation of $Dgm(d_{\mu,m})$



Section 3: Data Sparsification

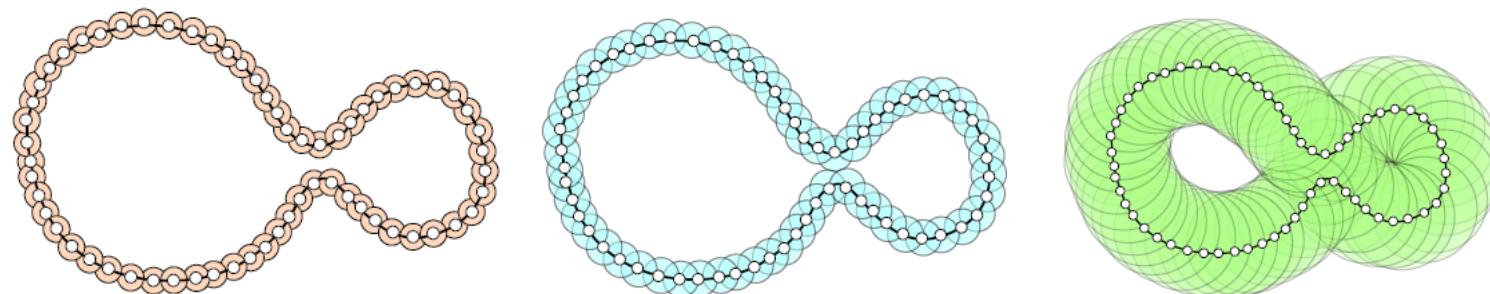
Rips Filtration

Rips Filtration

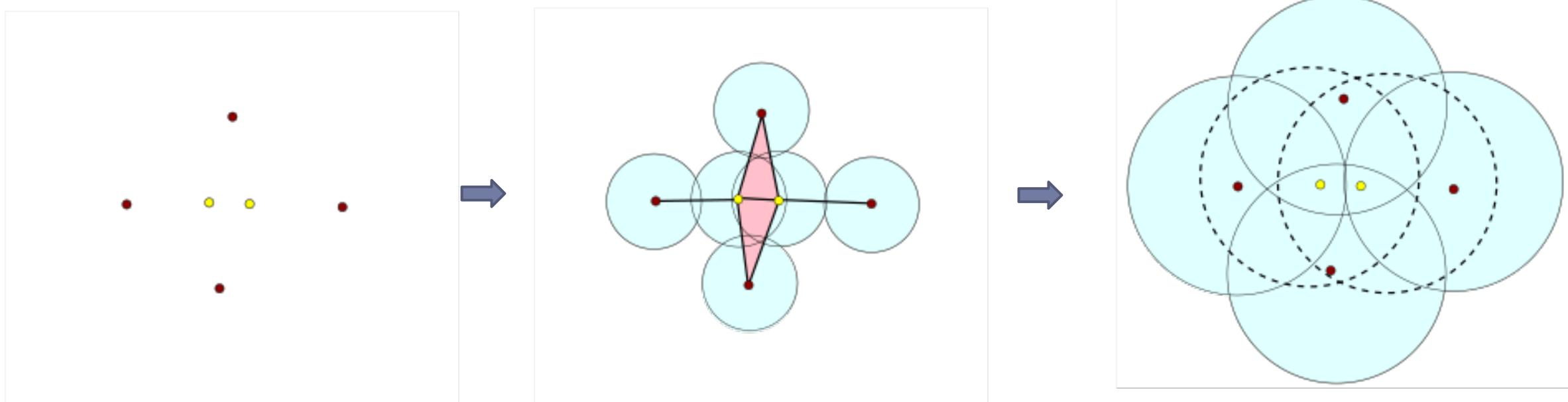


Rips Filtration

- ▶ Size becomes huge quickly
- ▶ But not all simplices are needed, especially at large scales!

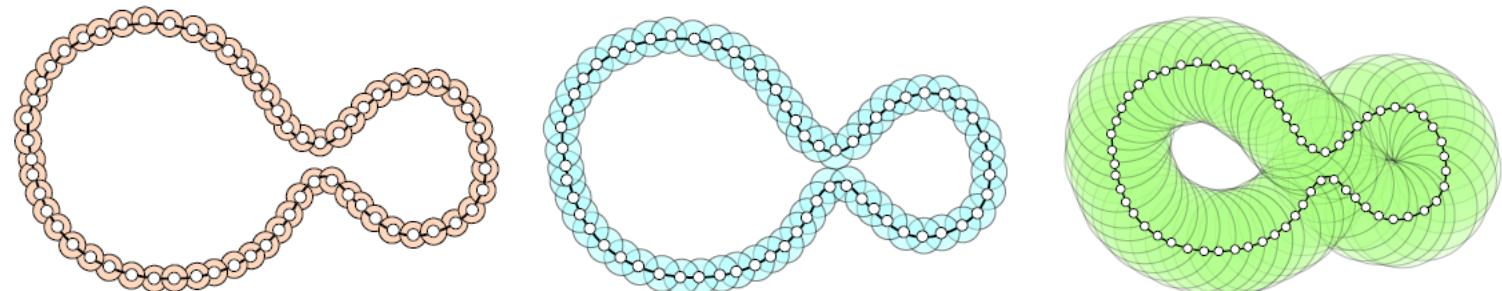


Rips Filtration



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
- ▶ How to achieve that?
 - ▶ Intuitively, use a $\frac{\delta}{c}$ -sparse sample to produce union of balls at scale $r = \delta$
 - ▶ [Sheehy, DCG 2013]



How to define such 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 $p \in P$, $d(p, Q) \leq \varepsilon$, i.e., $d_H(P, Q) \leq \varepsilon$
 - ▶ (sparsity-condition): Q is a ε -sparse, i.e., for every $q \neq q' \in Q$, $d(q, q') \geq \varepsilon$
- ▶ 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

Net-tower

- ▶ Given input point set $P \subset R^d$
- ▶ Let $\{p_1, p_2, \dots, p_n\}$ be the sequence of points obtained via *farthest point sampling procedure*

Net-tower

- ▶ Given input point set $P \subset R^d$
- ▶ Let $\{p_1, p_2, \dots, p_n\}$ be the sequence of points obtained via *farthest point sampling procedure*
 - ▶ Pick arbitrary point $p_1 \in P$ and set $P_1 = \{p_1\}$

Net-tower

- ▶ Given input point set $P \subset R^d$
- ▶ Let $\{p_1, p_2, \dots, p_n\}$ 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 $p_i = \operatorname{argmax}_{p \in P \setminus P_{i-1}} d(p, P_{i-1})$. Set $P_i = P_{i-1} \cup \{p_i\}$

Net-tower

- ▶ Given input point set $P \subset R^d$
- ▶ Let $\{p_1, p_2, \dots, p_n\}$ 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 $p_i = \operatorname{argmax}_{p \in P \setminus P_{i-1}} d(p, P_{i-1})$. Set $P_i = P_{i-1} \cup \{p_i\}$
 - ▶ Easy to see that $P = P_n \supset P_{n-1} \supset \dots \supset P_2 \supset P_1$

Net-tower

- ▶ Given input point set $P \subset R^d$
- ▶ Let $\{p_1, p_2, \dots, p_n\}$ 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 $p_i = \operatorname{argmax}_{p \in P \setminus P_{i-1}} d(p, P_{i-1})$. Set $P_i = P_{i-1} \cup \{p_i\}$
 - ▶ Easy to see that $P = P_n \supset P_{n-1} \supset \dots \supset P_2 \supset P_1$
- ▶ Exit-time of $p = p_i$ is set to be $t_{p_i} := d(p_i, P_{i-1})$

Net-tower

- ▶ Given input point set $P \subset R^d$
- ▶ Let $\{p_1, p_2, \dots, p_n\}$ be the sequence of points obtained via *farthest point sampling procedure*

Net-tower

- ▶ Given input point set $P \subset R^d$
- ▶ Let $\{p_1, p_2, \dots, p_n\}$ be the sequence of points obtained via *farthest point sampling procedure*
 - ▶ Pick arbitrary point $p_1 \in P$ and set $P_1 = \{p_1\}$

Net-tower

- ▶ Given input point set $P \subset R^d$
- ▶ Let $\{p_1, p_2, \dots, p_n\}$ 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 $p_i = \operatorname{argmax}_{p \in P \setminus P_{i-1}} d(p, P_{i-1})$. Set $P_i = P_{i-1} \cup \{p_i\}$

Net-tower

- ▶ Given input point set $P \subset R^d$
- ▶ Let $\{p_1, p_2, \dots, p_n\}$ 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 $p_i = \operatorname{argmax}_{p \in P \setminus P_{i-1}} d(p, P_{i-1})$. Set $P_i = P_{i-1} \cup \{p_i\}$
 - ▶ Easy to see that $P = P_n \supset P_{n-1} \supset \dots \supset P_2 \supset P_1$

Net-tower

- ▶ Given input point set $P \subset R^d$
- ▶ Let $\{p_1, p_2, \dots, p_n\}$ 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 $p_i = \operatorname{argmax}_{p \in P \setminus P_{i-1}} d(p, P_{i-1})$. Set $P_i = P_{i-1} \cup \{p_i\}$
 - ▶ Easy to see that $P = P_n \supset P_{n-1} \supset \dots \supset P_2 \supset P_1$
- ▶ Exit-time of $p = p_i$ is set to be $t_{p_i} := d(p_i, P_{i-1})$

Net-tower

- ▶ Given input point set $P \subset R^d$
- ▶ Let $\{p_1, p_2, \dots, p_n\}$ 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 $p_i = \operatorname{argmax}_{p \in P \setminus P_{i-1}} d(p, P_{i-1})$. Set $P_i = P_{i-1} \cup \{p_i\}$
 - ▶ Easy to see that $P = P_n \supset P_{n-1} \supset \dots \supset P_2 \supset P_1$
- ▶ Exit-time of $p = p_i$ is set to be $t_{p_i} := d(p_i, P_{i-1})$
- ▶ Each P_i is a t_{p_i} -net of P

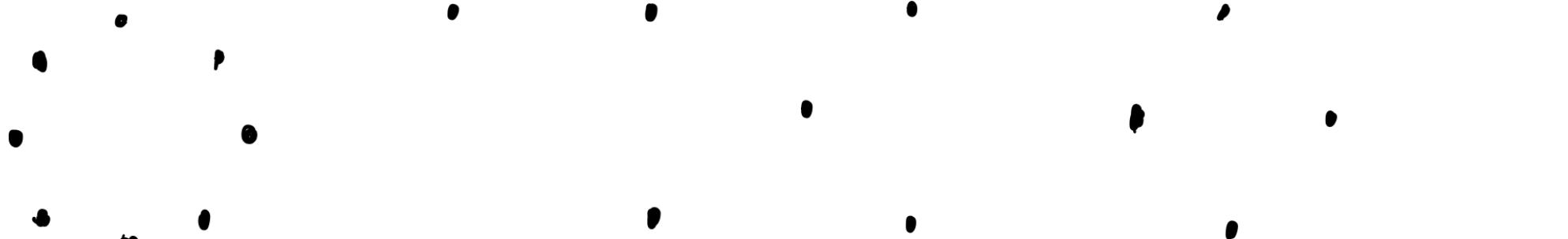
Net-tower

- ▶ Given input point set $P \subset \mathbb{R}^d$
- ▶ Let $\{p_1, p_2, \dots, p_n\}$ 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 $p_i = \operatorname{argmax}_{p \in P \setminus P_{i-1}} d(p, P_{i-1})$. Set $P_i = P_{i-1} \cup \{p_i\}$
 - ▶ Easy to see that $P = P_n \supset P_{n-1} \supset \dots \supset P_2 \supset P_1$
- ▶ Exit-time of $p = p_i$ is set to be $t_{p_i} := d(p_i, P_{i-1})$
- ▶ Construct two families of γ -nets:

Open net-tower $\mathcal{N} = \{N_\gamma\}_{\gamma \in \mathbb{R}}$ where $N_\gamma := \{p \in P \mid t_p > \gamma\}$.

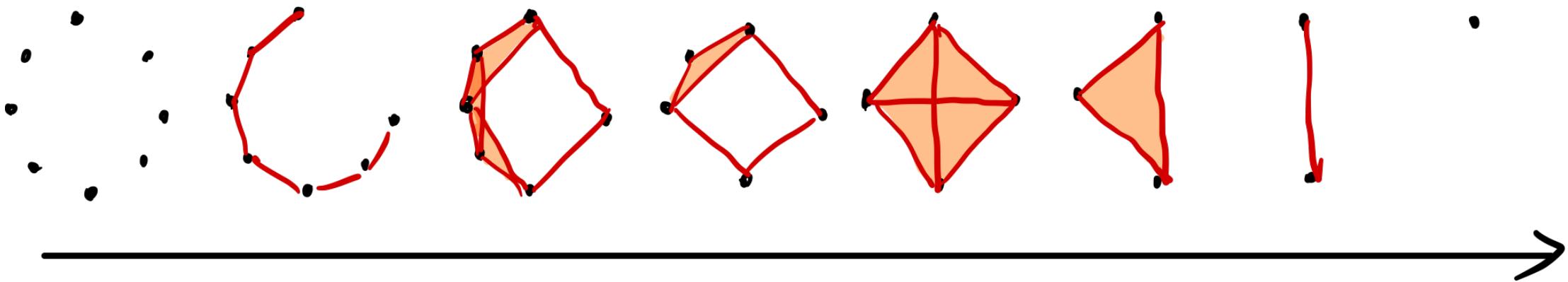
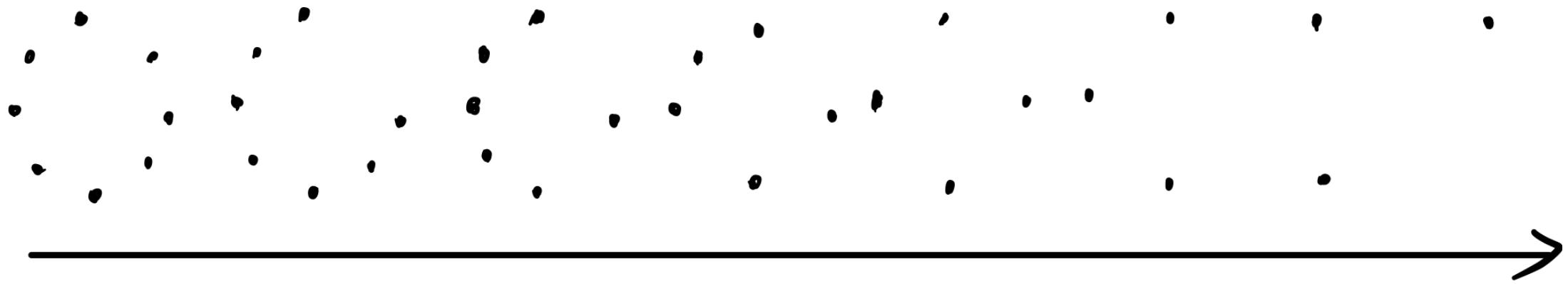
Closed net-tower $\overline{\mathcal{N}} = \{\overline{N}_\gamma\}_{\gamma \in \mathbb{R}}$ where $\overline{N}_\gamma := \{p \in P \mid t_p \geq \gamma\}$.

Example

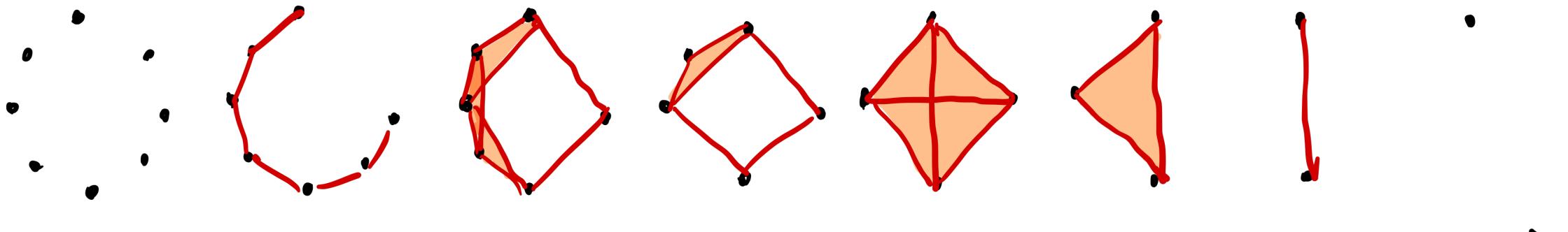


P_1 P_2 P_3 P_4

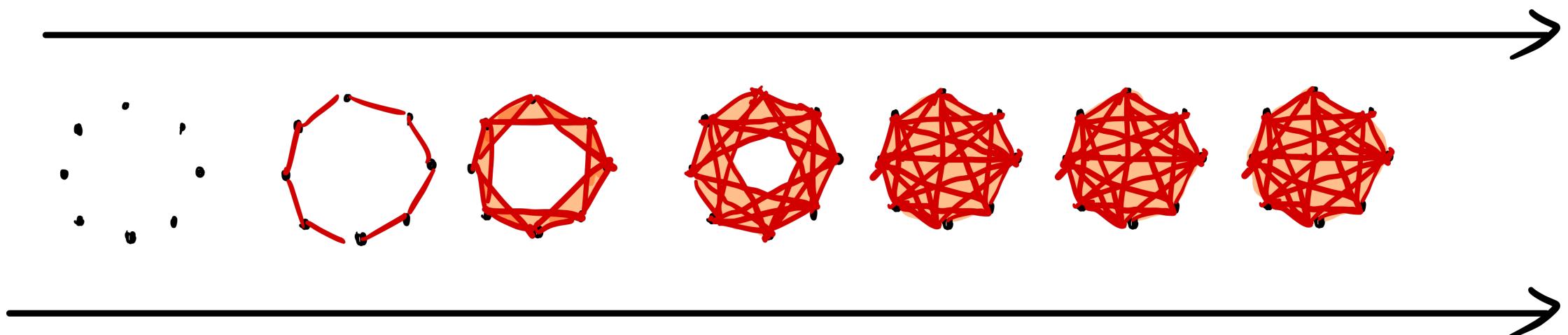
P_5 P_6 P_7 P_8



Sparse Rips filtration



Rips filtration



Approximation via simplicial tower

- Given a set of points $P \subset \mathbb{R}^d$, $\alpha > 0$ and some $0 < \varepsilon < 1$, consider the following filtration, which is a subsequence of standard Rips filtration:

$$\mathbb{VR}^\alpha(P) \hookrightarrow \mathbb{VR}^{\alpha(1+\varepsilon)}(P) \hookrightarrow \mathbb{VR}^{\alpha(1+\varepsilon)^2}(P) \hookrightarrow \dots \hookrightarrow \mathbb{VR}^{\alpha(1+\varepsilon)^m}(P).$$

- We will aim to approximate the above sequence via a sequence of sparsified simplicial complexes connected by simplicial maps (not inclusions)

Nets, and induced Rips complexes

- ▶ Set $P_0 := P$. Build a sequence of point sets P_1, \dots, P_m such that
 - ▶ P_{k+1} is a $\left(\frac{\alpha\varepsilon}{2}\right)(1 + \varepsilon)^{k-1}$ -net of P_k
 - ▶ Terminates when P_m is of constant size.

Nets, and induced Rips complexes

- ▶ Set $P_0 := P$. Build a sequence of point sets P_1, \dots, P_m such that
 - ▶ P_{k+1} is a $\left(\frac{\alpha\varepsilon}{2}\right)(1 + \varepsilon)^{k-1}$ -net of P_k
 - ▶ Terminates when P_m is of constant size.
- ▶ Vertex map $\pi_k: P_k \rightarrow P_{k+1}$
 - ▶ where $\pi_k(p) = \operatorname{argmin}_{q \in P_{k+1}} d(p, q)$

Nets, and induced Rips complexes

- ▶ Set $P_0 := P$. Build a sequence of point sets P_1, \dots, P_m such that
 - ▶ P_{k+1} is a $\left(\frac{\alpha\varepsilon}{2}\right)(1 + \varepsilon)^{k-1}$ -net of P_k
 - ▶ Terminates when P_m is of constant size.
- ▶ Vertex map $\pi_k: P_k \rightarrow P_{k+1}$
 - ▶ where $\pi_k(p) = \operatorname{argmin}_{q \in P_{k+1}} d(p, q)$
- ▶ Claim:
 - ▶ π_k induces a simplicial map $\pi_k : \mathbb{VR}^{\alpha(1+\varepsilon)^k}(P_k) \rightarrow \mathbb{VR}^{\alpha(1+\varepsilon)^{k+1}}(P_{k+1})$

Nets, and induced Rips complexes

- ▶ Set $P_0 := P$. Build a sequence of point sets P_1, \dots, P_m such that
 - ▶ P_{k+1} is a $\left(\frac{\alpha\varepsilon}{2}\right)(1+\varepsilon)^{k-1}$ -net of P_k
 - ▶ Terminates when P_m is of constant size.
- ▶ Vertex map $\pi_k: P_k \rightarrow P_{k+1}$
 - ▶ where $\pi_k(p) = \operatorname{argmin}_{q \in P_{k+1}} d(p, q)$
- ▶ Claim:
 - ▶ π_k induces a simplicial map $\pi_k : \mathbb{VR}^{\alpha(1+\varepsilon)^k}(P_k) \rightarrow \mathbb{VR}^{\alpha(1+\varepsilon)^{k+1}}(P_{k+1})$
- ▶ This thus gives rise to the following simplicial tower:

$$\widehat{\mathcal{S}} : \mathbb{VR}^\alpha(P_0) \xrightarrow{\pi_0} \mathbb{VR}^{\alpha(1+\varepsilon)}(P_1) \xrightarrow{\pi_1} \dots \xrightarrow{\pi_{m-1}} \mathbb{VR}^{\alpha(1+\varepsilon)^m}(P_m).$$

Theoretical Guarantee

Claim 6.3. For any fixed $\alpha \geq 0$, $\varepsilon \geq 0$, and any integer $k \geq 0$, each triangle in the following diagram commutes at the homology level:

$$\begin{array}{ccc} \mathbb{VR}^{\alpha(1+\varepsilon)^k}(P_0) & \xrightarrow{i_k} & \mathbb{VR}^{\alpha(1+\varepsilon)^{k+1}}(P_0) \\ j_k \uparrow & \searrow \widehat{\pi}_k & j_{k+1} \uparrow \\ \mathbb{VR}^{\alpha(1+\varepsilon)^k}(P_k) & \xrightarrow{\pi_k} & \mathbb{VR}^{\alpha(1+\varepsilon)^{k+1}}(P_{k+1}) \end{array}$$

Here, the maps i_k s and j_k s are canonical inclusions.

Theoretical Guarantee

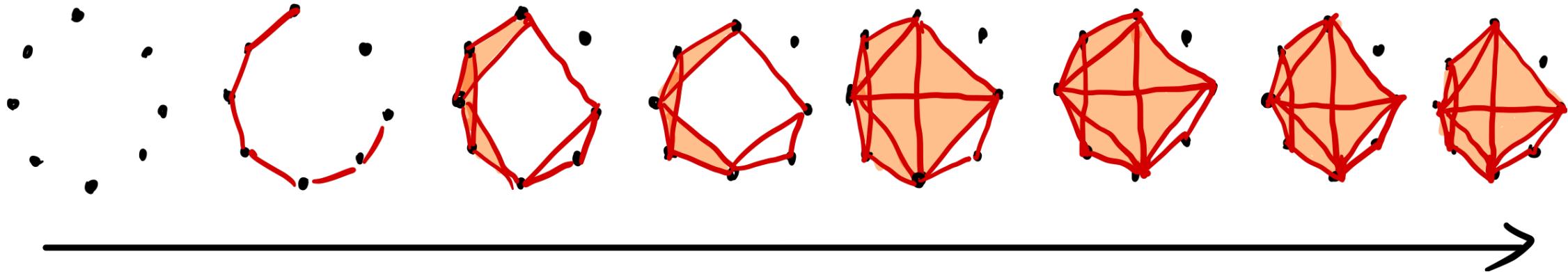
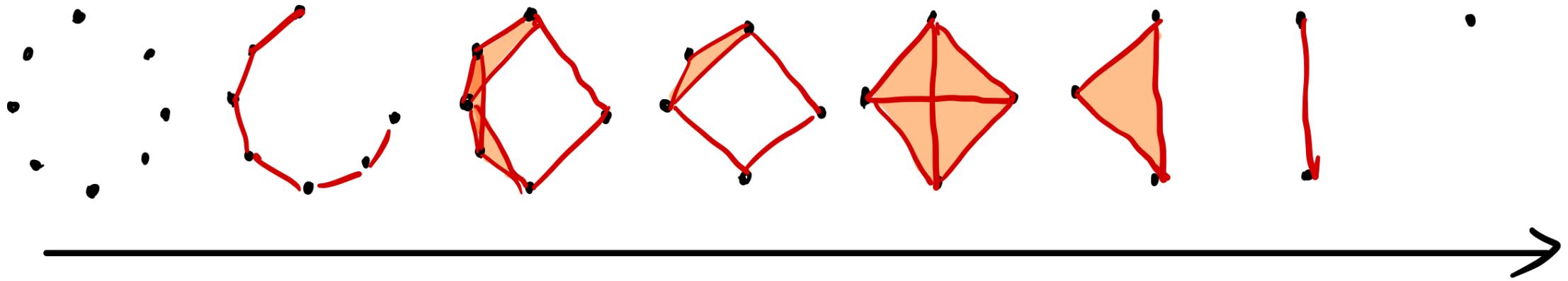
Claim 6.3. For any fixed $\alpha \geq 0$, $\varepsilon \geq 0$, and any integer $k \geq 0$, each triangle in the following diagram commutes at the homology level:

$$\begin{array}{ccc} \mathbb{VR}^{\alpha(1+\varepsilon)^k}(P_0) & \xrightarrow{i_k} & \mathbb{VR}^{\alpha(1+\varepsilon)^{k+1}}(P_0) \\ j_k \uparrow & \searrow \widehat{\pi}_k & j_{k+1} \uparrow \\ \mathbb{VR}^{\alpha(1+\varepsilon)^k}(P_k) & \xrightarrow{\pi_k} & \mathbb{VR}^{\alpha(1+\varepsilon)^{k+1}}(P_{k+1}) \end{array}$$

Here, the maps i_k s and j_k s are canonical inclusions.

Theorem 6.8. Given a set of n points $P \subset \mathbb{R}^d$, we can $3\log(1 + \varepsilon)$ -approximate the persistence diagram of the discrete Rips filtration in Eqn. (6.12) by that of the filtration in Eqn. (6.13) at the log-scale. The number of k -simplices that ever appear in the filtration in Eqn. (6.13) is $O((\frac{1}{\varepsilon})^{O(kd)} n)$.

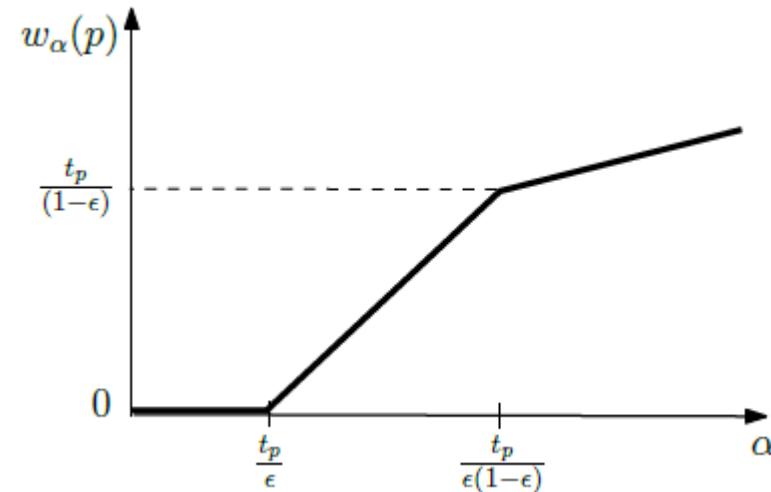
Another approach with inclusion maps



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:

- ▶ $\hat{d}_\alpha(p, q) = d(p, q) + w_p(\alpha) + w_q(\alpha)$

Sparse Rips complexes

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 $\{\bar{N}_\gamma\}$ for P as introduced above, the *open sparse-Rips complex at scale α* is defined as

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

Sparse Rips complexes

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

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

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

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

Sparse Rips complexes

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

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

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

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

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

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 $\{\bar{N}_\gamma\}$ induced by exit-times $\{t_p\}_{p \in P}$, let $\mathcal{S}(P) = \{\mathbb{S}^\alpha\}$ 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 $\text{Dgm}_k \mathcal{S}(P)$ is a $\log \frac{1}{1-\varepsilon}$ -approximation of $\text{Dgm}_k \mathcal{R}(P)$ 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 $\Theta((\frac{1}{\varepsilon})^{kd} n)$.*

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

Another approach for data sparsification

FIN