Persistent Homology and its stability

김지수 (Jisu KIM)

통계이론세미나 - 위상구조의 통계적 추정, 2023 가을학기

When analyzing data, we prefer robust features where features of the underlying manifold can be inferred from features of finite samples. As I introduced before, homology is counting holes. But there is a problem of using homology to data: that homology of finite sample is very differnt from homology of underlying manifold. See Figure .

We first recall the homology.

Definition. Let K be a simplicial complex, $k \geq 0$ be a nonnegative integer, and G be an abelian group. The space of k-chains on K, $C_k(K;G)$, is the set whose elements are a finite formal sum of k-simplices of K with coefficients from G, i.e.,

$$C_k(K;G) = \left\{ \sum_i n_i \sigma_i : n_i \in G, \sigma_i \in K_k \right\},$$

where $K_k \subset K$ is the set of k-simplices of K. We write $C_k(K)$ if the coefficient group G is understood from the context.

For an integer $k \leq -1$, we define $C_k(K) = 0$ for convenience.

Remark. Typical examples of G are $G = \mathbb{Z}$ and $G = \mathbb{Z}_2 = \mathbb{Z}/2\mathbb{Z}$. For $G = \mathbb{Z}_2$, $C_k(K;\mathbb{Z}_2)$ becomes a vector space.

Remark. $C_k(K;G)$ has an abelian group structure as for $\sum_i n_i\sigma_i, \sum_i n_i'\sigma_i \in C_k(K;G)$,

$$\left(\sum_i n_i\sigma_i\right) + \left(\sum_i n_i'\sigma_i\right) \coloneqq \sum_i (n_i + n_i')\sigma_i.$$

When G is a field, $C_k(K;G)$ has a natural vector space structure as for $\sum_i n_i\sigma_i \in C_k(K;G)$ and $\lambda \in G$,

$$\lambda \cdot \left(\sum_i n_i\sigma_i\right) = \sum_i (\lambda \cdot n_i)\sigma_i.$$

To relate chain groups of different dimensions, we define the boundary map as sending each -simplex to the sum of its $(k-1)$-dimensional faces. We write $\sigma = [v_0, \ldots, v_k]$ for an ordered simplex, i.e., $[v_0, v_1] = -[v_1, v_0]$.

Definition. A boundary map $\partial_k : C_k(K) \to C_{k-1}(K)$ is defined for each simplex as (see Figure)

$$\partial_k [v_0, \ldots, v_k] = \sum_{j=0}^{k} (-1)^j [v_0, \ldots, \hat{v}_j, \ldots, v_k],$$

where $[v_0, \ldots, \hat{v}_j, \ldots, v_k] = [v_0, \ldots, v_{j-1}, v_{j+1}, \ldots, v_k] \in K_{k-1}$, i.e., $\hat{v}_j$ means that v_j is omitted. The definition is extended to entire k-chain as

$$\partial_k \left(\sum_i n_i\sigma_i\right) = \sum_i n_i \partial_k \sigma_i.$$

Remark. ∂_k satisfies that for $c, c' \in C_k(K)$, $\partial_k(c + c') = \partial_k c + \partial_k c'$, so $\partial_k : C_k(K) \to C_{k-1}(K)$ is a homomorphism.

Lemma ([4, Lemma 2.1])**.** $\partial_{k-1} \circ \partial_k = 0$.

Definition. Cycles and boundaries

(a) A k-cycle group $Z_k = Z_k(K)$ is the k-cycle whose boundary is 0, i.e.,

$$Z_k(K) = \ker \partial_k = \{c \in C_k(K) : \partial_k c = 0\}.$$

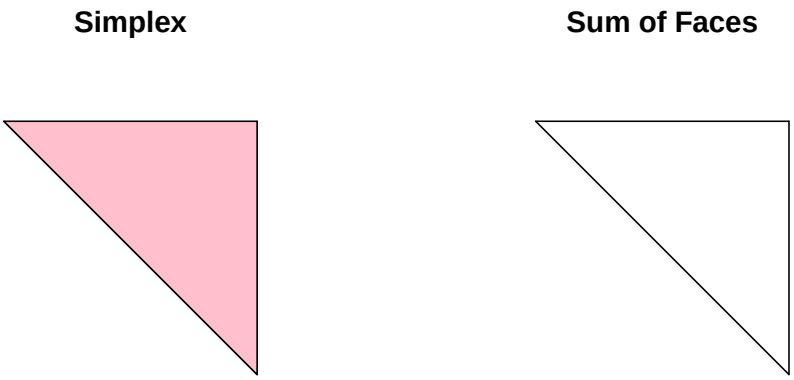


Figure 1: Boundary map.

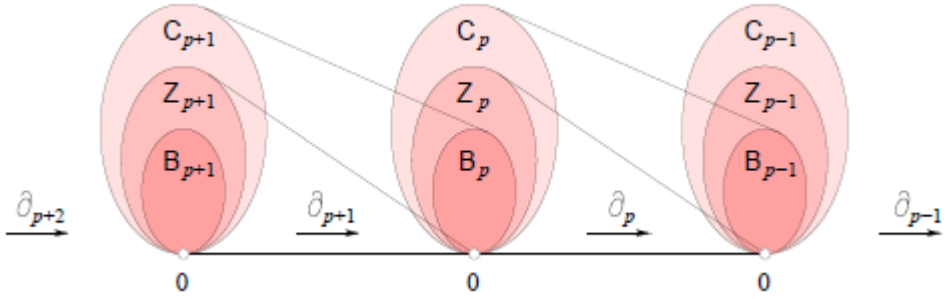


Figure 2: [3, Figure IV.1] Interleaving relations between cycle groups and boundary groups via boundary map.

(b) A k-boundary group $B_k = B_k(K)$ is the k-cycle that is a boundary of $(k+1)$-chain,

$$B_k(K) = \mathrm{im}\partial_{k+1} = \{\partial_{k+1} d \in C_k(K) : d \in C_{k+1}(K)\}.$$

Then the above Lemma implies that $B_k(K)$, $Z_k(K)$, $C_k(K)$ are interleaved as subgroups (see Figure):

$$B_k(K) \subset Z_k(K) \subset C_k(K).$$

Definition. The k-th homology group is the k-th cycle group modulo the k-th boundary group,

$$H_k = H_k(K) \coloneqq Z_k(K)/B_k(K).$$

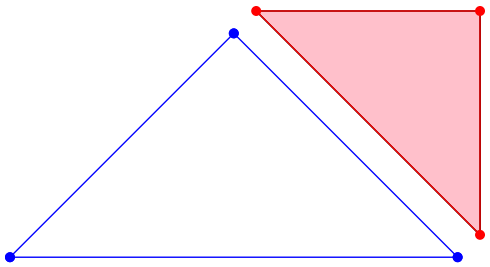
The k-th Betti number is the rank of this group, $\beta_k = \mathrm{rank} H_k$.

Example. Suppose K is given as the right of Figure , and use $G = \mathbb{Z}$. Then for $k = 1$, its cycle group, boundary group, homology group, and betti number is computed as in Figure .

Persistent homology is a multiscale approach to represent topological features.

A *filtration* $\mathcal{F}$ is a collection of objects (subcomplexes or subsets) approximating the data points at different resolutions, formally defined as follows.

Definition (Filtration)**.** A *filtration* $\mathcal{F}$ is a collection of increasing objects:



- $Z_1(K) = \ker \partial_1 = \mathbb{Z}^2 =<$  ,  $>$
- $B_1(K) = \mathrm{im}\partial_2 = \mathbb{Z} =<$  $>$
- $H_1(K) = Z_1(K)/B_1(K) = \mathbb{Z} =<$  $>$, $\beta_1(K) = 1$

Figure 3: Homology example.

- Let K be a simplicial complex. A (subcomplexes) *filtration* $\mathcal{F} = \{\mathcal{F}_a \subset K\}_{a\in\mathbb{R}}$ is a collection of subcomplexes of K such that $a \leq b$ implies that $\mathcal{F}_a \subset \mathcal{F}_b$.
- A (subsets) *filtration* $\mathcal{F} = \{\mathcal{F}_a \subset \mathbb{X}\}_{a\in\mathbb{R}}$ is a collection of subsets of a topological space $\mathbb{X}$ such that $a \leq b$ implies that $\mathcal{F}_a \subset \mathcal{F}_b$.

A typical way of setting the filtration is through a real-valued function.

Definition. For a simplicial complex K, a real function $f : K \to \mathbb{R}$ is monotonic if $f(\sigma) \leq f(\tau)$ whenever σ is a face of τ.

For a real monotonic function $f : K \to \mathbb{R}$, if we let $\mathcal{F}_a \coloneqq f^{-1}(-\infty, a]$, then the monotonicity implies that $\mathcal{F}_a$ is a subcomplex of K and $\mathcal{F}_a \subset \mathcal{F}_b$ whenever $a \leq b$, so $\mathcal{F} = \{\mathcal{F}_a \subset K\}_{a\in\mathbb{R}}$ is a subcomplexes filtration. Similarly, for a real function $f : \mathbb{X} \to \mathbb{R}$ on a topological space $\mathbb{X}$ (not necessarily continuous), if we let $\mathcal{F}_a \coloneqq f^{-1}(-\infty, a]$, then $\mathcal{F} = \{\mathcal{F}_a \subset K\}_{a\in\mathbb{R}}$ is a subsets filtration. This filtration $\mathcal{F}$ is called a sublevel filtration (of f).

Remark. We also consider a superlevel filtration $\left\{f^{-1}[a, \infty)\right\}_{a\in\mathbb{R}}$, in particular when f is a density function. However, a superlevel filtration $\left\{f^{-1}[a, \infty)\right\}_{a\in\mathbb{R}}$ is equivalent to a sublevel filtration $\left\{(-f)^{-1}(-\infty, a]\right\}_{a\in\mathbb{R}}$, so for this lecture note we just consider the sublevel filtrations.

For a filtration $\mathcal{F}$ and for each $k \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$, the associated persistent homology $PH_k\mathcal{F}$ is an ordered collection of k-th dimensional homologies, one for each element of $\mathcal{F}$.

Definition (Persistent Homology)**.** Let $\mathcal{F}$ be a filtration and let $k \in \mathbb{N}_0$. The associated k-th *persistent homology* $PH_k\mathcal{F}$ is a collection of groups $\{H_k\mathcal{F}_a\}_{a\in\mathbb{R}}$ equipped with homomorphisms $\{\imath_k^{a,b}\}_{a\leq b}$, where $H_k\mathcal{F}_a$ is the k-th dimensional homology group of $\mathcal{F}_a$ and $\imath_k^{a,b} : H_k\mathcal{F}_a \to H_k\mathcal{F}_b$ is the homomorphism induced by the inclusion $\mathcal{F}_a \subset \mathcal{F}_b$. Write $H_k^{a,b} \coloneqq \mathrm{im}(\imath_k^{a,b})$. The corresponding k-th *persistent Betti numbers* are the ranks of these groups, $\beta_k^{a,b} = \mathrm{rank}H_k^{a,b}$.

The persistent homology groups $H_k^{a,b}$ consist of homology classes of $\mathcal{F}_a$ that are still alive at $\mathcal{F}_b$, or moreformally, $H_k^{a,b} = Z_k(\mathcal{F}_a)/(B_k(\mathcal{F}_b) \cap Z_k(\mathcal{F}_a))$.

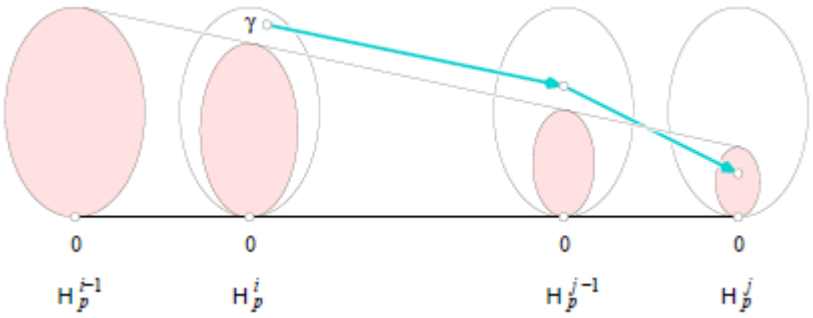


Figure 4: [3, Figure VII.2] The class is γ is born at i since it does not lie in the (shaded) image of H_k^{i-1}. Furthermore, γ dies entering j since this is the first time its image merges into the image of H_k^{i-1}.

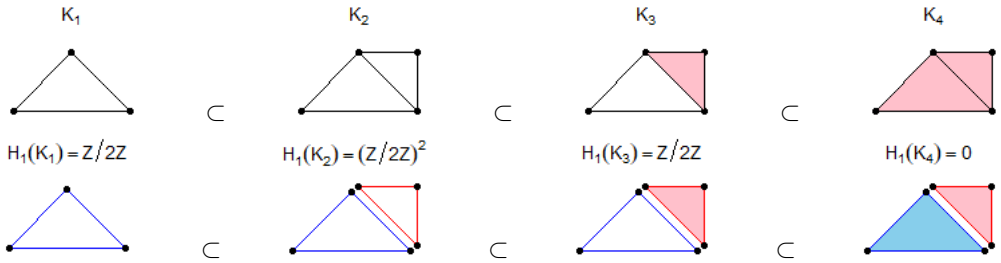


Figure 5: Persistent Homology from subcomplexes of a simplicial complex.

For the k-th persistent homology $PH_k\mathcal{F}$, the set of filtration levels at which a specific homology appears is always an interval $[b, d) \subset [-\infty, \infty]$, i.e. a specific homology is formed at some filtration value b and dies when the inside hole is filled at another value $d > b$. To be more concrete about the classes counted by the persistent homology groups, let γ be a class in $H_k(\mathcal{F}_b)$. We say it is born at $b \in \mathbb{R}$ if $\gamma \notin H_k^{a,b}$ for all $a < b$. Furthermore, if γ is born at b then it dies entering d if it merges with an older class as we go from $\mathcal{F}_c$ to $\mathcal{F}_d$ for any $c \in [b, d)$, i.e., for any $c \in [b, d)$, there exists $a < b$ such that $\imath_k^{b,c}(\gamma) \notin H_k^{a,c}$ but $\imath_k^{b,d}(\gamma) \in H_k^{a,d}$. See Figure . In this sense,

We visualize the collection of persistent Betti numbers by drawing points in two dimensions.

Definition (Persistence Diagram)**.** Let $\mathcal{F}$ be a filtration and let $k \in \mathbb{N}_0$. The corresponding k-th persistence diagram $Dgm_k(\mathcal{F})$ is a finite multiset of $(\mathbb{R} \cup \{\infty\})^2$, consisting of all pairs (b, d) where $[b, d)$ is the interval of filtration values for which a specific homology appears in $PH_k\mathcal{F}$. b is called a birth time and d is called a death time.

When does the persistence diagram "fully" represent all the topological information in the persistent homology? There are two sufficient conditions [1, Theorem 2.8]:

- if $\{H_k\mathcal{F}_a\}_{a\in\mathbb{R}}$ changes only finite times, i.e., $\{\imath_k^{a,b} : H_k\mathcal{F}_a \to H_k\mathcal{F}_b\}_{a\leq b}$ is not an isomorphism for finitely many $a_1, \ldots, a_m$.
- or if for all $a \in \mathbb{R}$, $\text{rank}(H_k\mathcal{F}_a)$ is finite.

And this covers all the practical cases.

Example. See Figure .

Example. See Figure . Suppose we want to find a loop structure of a circle, from 20 data points on a circle. We attach disks of radius r to each data point, and increase the radius r from 0 to ∞. When $r = 0.5$, the collection of disks form a loop, and this is the birth time of the loop. When $r = 1$, the inside hole is filled, and this is the death time of the loop. Then we collect birth time and death time of all possible loops, and this is the persistent homology / persistence diagram.

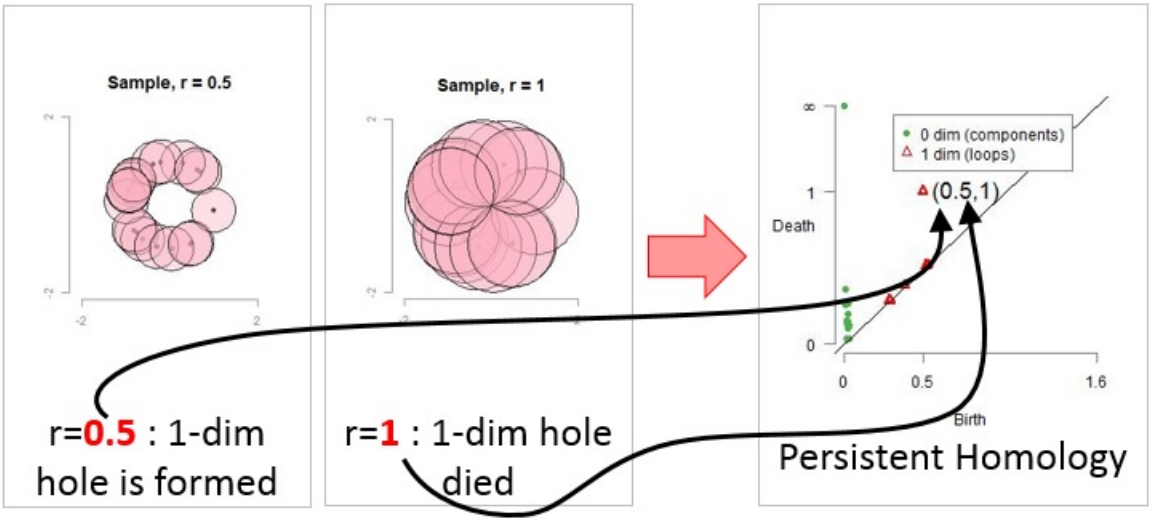


Figure 6: Persistent Homology from Cech filtration.

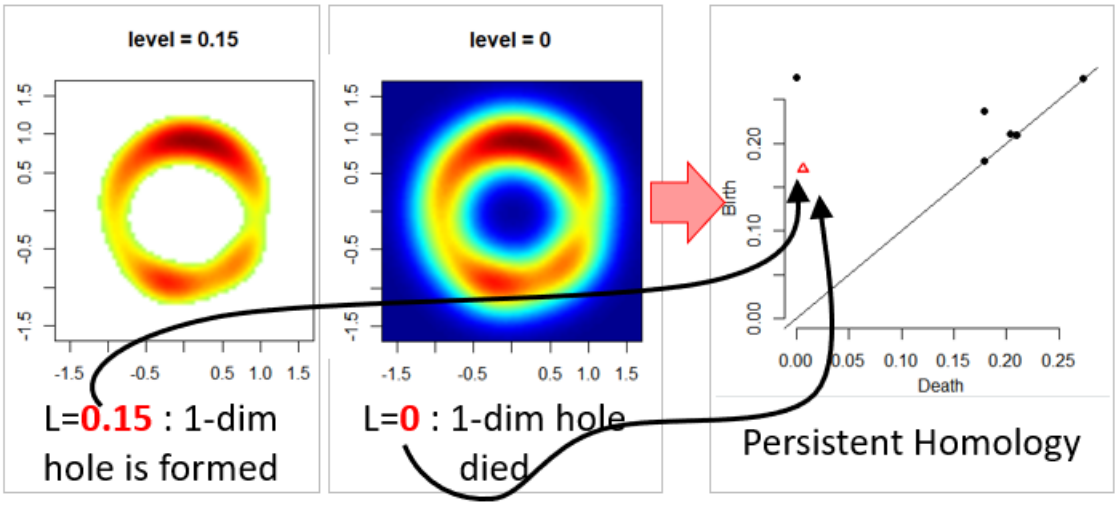


Figure 7: Persistent Homology from Kde filtration.

Example. See Figure . Suppose we consider superlevel sets of the kernel density estimator. We decrease level L from ∞ to 0. When $L = 0.15$, you can see that 1-dim hole is formed, and this is the birth time of this loop. And as you decrease L, the inside hole becomes smaller, and when $L = 0$, then inside hole is filled, and this is the death time of this loop. Then we collect birth time and death time of all possible loops, and this is the persistent homology / persistence diagram.

Stability Theorem

This section is mostly from [1].

To impose stability, we first endow the space of persistence diagrams with a metric. The most fundamental one is the bottleneck distance.

Definition. The bottleneck distance between two persistence diagrams $Dgm_k(\mathcal{F})$ and $Dgm_k(\mathcal{G})$ is defined by

$$d_B(Dgm_k(\mathcal{F}), Dgm_k(\mathcal{G})) = \inf_{\gamma\in\Gamma} \sup_{p\in Dgm_k(\mathcal{F})} \|p - \gamma(p)\|_\infty,$$

where the set Γ consists of all the bijections $\gamma : Dgm_k(\mathcal{F}) \cup Diag \to Dgm_k(\mathcal{G}) \cup Diag$, and $Diag$ is the diagonal $\{(x, x) : x \in \mathbb{R}\} \subset \mathbb{R}^2$ with infinite multiplicity.

The bottleneck distance imposes a metric structure on the space of persistence diagrams, which leads to the persistence stability theorem, i.e., small perturbations in the data implies at most small changes in the persistence diagrams in terms of the bottleneck distance.

For the stability theorem, we consider the case that the coefficient group G is a field so that the homology groups $H_k\mathcal{F}_a$ are vector spaces. Recall that the k-th dimensional persistent homology is a collection of groups $\{H_k\mathcal{F}_a\}_{a\in\mathbb{R}}$ equipped with homomorphisms $\{\imath_k^{a,b}\}_{a\le b}$.

Definition. A *persistence module* $\mathcal{PF}$ for a filtration $\mathcal{F}$ is a family $\{F_a\}_{a\in\mathbb{R}}$ of vector spaces, together with a family $\{\jmath^{a,b}: F_a \to F_b\}_{a\le b}$ of homomorphisms such that: $\forall a \le b \le c$, $\jmath^{a,c} = \jmath^{b,c}\circ\jmath^{a,b}$ and $\jmath^{a,a} = id_{F_a}$.

In this lecture note, for a function $f:\mathbb{X}\to\mathbb{R}$ defined on a metric space $\mathbb{X}$, let $\mathcal{P}(f)$ be the persistence module induced from sublevel filtrations $\left\{H_k f^{-1}(-\infty,a]\right\}_{a\in\mathbb{R}}$ (and inclusion homomorphism maps). For a metric space $\mathcal{X}$, let $\mathcal{PR}(\mathcal{X})$ be the persistence module induced from Vietoris-Rips complexes $\{H_k\mathrm{Rips}(\mathcal{X},r)\}_{r\in\mathbb{R}}$. For a metric space $\mathbb{X}$ and $\mathcal{X}\subset\mathbb{X}$, let $\mathcal{PC}_{\mathbb{X}}(\mathcal{X})$ be the persistence module induced from Čech complexes $\left\{H_k\check{\mathrm{C}}\mathrm{ech}_{\mathbb{X}}(\mathcal{X},r)\right\}_{r\in\mathbb{R}}$.

We will impose a standard regularity condition for the persistence module $\mathcal{PF}$, which is *tameness*.

Definition ([1, Section 3.8])**.** Let $\mathcal{PF}=\{F_a\}_{a\in\mathbb{R}}$ be a persistence module. A persistence module $\mathcal{PF}$ is q-*tame* if the image $im(\jmath^{a,b})$ of the homomorphism $\jmath^{a,b}: F_a\to F_b$ is of finite rank for all $a<b$.

For the persistence module induced from the sublevel sets $\mathcal{P}(f)=\left\{f^{-1}(-\infty,a]\right\}_{a\in\mathbb{R}}$, a sufficient condition for the q-tame is that the image of $f:\mathbb{R}^d\to\mathbb{R}$ is lower bounded and proper.

Proposition ([1, Corollary 3.34])**.** *Let $\mathbb{X}$ be an image of a locally finite simplicial complex (that is, there is a locally finite simplicial complex K such that $\mathbb{X}$ is a continuous image of $|K|$). Let $f:\mathbb{X}\to\mathbb{R}$ be a proper ($f^{-1}(C)$ for a compact set C is compact) continuous function that is bounded below. Then the persisence modules $\mathcal{P}(f)$ is q-tame.*

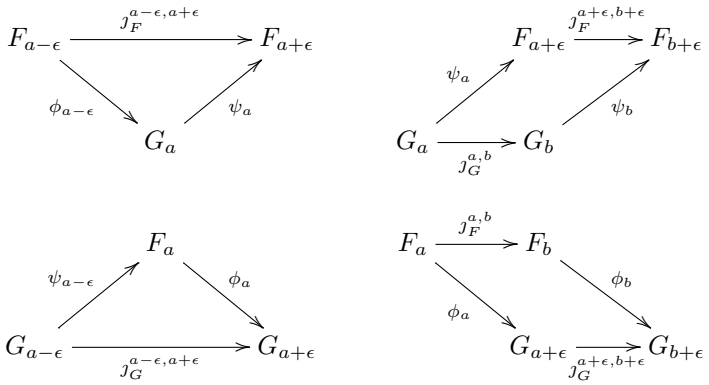
For persistence module induced from Čech complex $\check{\mathrm{C}}\mathrm{ech}_{\mathbb{X}}(\mathcal{X},r)$ or Vietoris-Rips complex $\mathrm{Rips}(\mathcal{X},r)$, a sufficient condition is that $\mathcal{X}$ is totally bounded:

Definition. A metric space X is *totally bounded* if for any $\epsilon>0$, there exists a finite set of points $x_1,\dots,x_n\in X$ that ϵ-approximates X, i.e. for all $x\in X$, there exists x_i such that $d(x,x_i)<\epsilon$.

Proposition ([2, Proposition 5.1])**.** *If $(\mathcal{X},d_{\mathcal{X}})$ is a totally bounded metric space, then the persistence modules $\mathcal{PR}(\mathcal{X})$ is q-tame. If $(\mathbb{X},d_{\mathbb{X}})$ is a totally bounded metric space and $\mathcal{X}\subset\mathbb{X}$, then the persistence modules $\mathcal{PC}_{\mathbb{X}}(\mathcal{X})$ is q-tame.*

For two functions $f,g:\mathbb{X}\to\mathbb{R}$ satisfying $\|f-g\|_\infty\le\epsilon$, their sublevel sets filtrations are nested as follows: $\forall a\in\mathbb{R}$, write $\mathcal{F}_a := f^{-1}(-\infty,a]$ and $\mathcal{G}_a := g^{-1}(-\infty,a]$, then $\mathcal{F}_a\subset\mathcal{G}_{a+\epsilon}$ and $\mathcal{G}_a\subset\mathcal{F}_{a+\epsilon}$. By letting $F_a=H_k(\mathcal{F}_a)$ and $G_a=H_k(\mathcal{G}_a)$, this induces the homomorphisms induced by the inclusions as $F_a\to G_{a+\epsilon}$ and $G_a\to F_{a+\epsilon}$. Also, the canonical inclusions $\mathcal{F}_a\subset\mathcal{F}_b$ and $\mathcal{G}_a\subset\mathcal{G}_b$ for $a\le b$ induces homomorphisms as $F_a\to F_b$ and $G_a\to G_b$. This homomorphisms relations can be extended as follows:

Definition ([1, Section 4.2])**.** Two persistence modules $\mathcal{PF}=\left\{\{F_a\}_{a\in\mathbb{R}},\{\jmath_F^{a,b}\}_{a\le b}\right\}$ and $\mathcal{PG}=\{G_a,\{\jmath_G^{a,b}\}_{a\le b}\}_{a\in\mathbb{R}}$ are said to be strongly ϵ-interleaved if there exist two families of homomorphisms $\{\phi_a: F_a\to G_{a+\epsilon}\}_{a\in\mathbb{R}}$ and $\{\psi_a: G_a\to F_{a+\epsilon}\}_{a\in\mathbb{R}}$ such that the following diagrams commute for all $a\le b$:



Definition ([1, Section 5.1])**.** The interleaving distance between two persistence modules $\mathcal{PF}$ and $\mathcal{PG}$ is defined as

$$d_I(\mathcal{PF},\mathcal{PG}) := \inf\{\epsilon>0 : \mathcal{PF}\text{ and }\mathcal{PG}\text{ are }\epsilon\text{-strongly interleaved.}\}$$

It is immediate that for two functions $f, g : \mathbb{X} \to \mathbb{R}$,

$$d_I(\mathcal{P}(f), \mathcal{P}(g)) \leq \|f - g\|_\infty .$$

Proposition ([2, Lemma 4.3, Corollary 4.10]). *1. For two metric spaces $\mathcal{X}, \mathcal{Y}$,*

$$d_I(\mathcal{PC}_{\mathcal{X}}(\mathcal{X}), \mathcal{PC}_{\mathcal{Y}}(\mathcal{Y})) \leq d_{GH}(\mathcal{X}, \mathcal{Y}),$$
$$d_I(\mathcal{PR}(\mathcal{X}), \mathcal{PR}(\mathcal{Y})) \leq d_{GH}(\mathcal{X}, \mathcal{Y}).$$

2. For two subsets $\mathcal{X}, \mathcal{Y} \subset \mathbb{X}$,

$$d_I(\mathcal{PC}_{\mathbb{X}}(\mathcal{X}), \mathcal{PC}_{\mathbb{X}}(\mathcal{Y})) \leq d_H(\mathcal{X}, \mathcal{Y}).$$

If two persistence modules are strongly interleaved, then their bottleneck distance are close, which is the strong stability theorem. Note that the bottleneck distance between filtrations are calculated on the persistence diagrams, so we can naturally consider the bottleneck distance to be defined on the persistence modules as well.

Theorem ([1, Theorem 5.23]). *Let $\mathcal{PF}$ and $\mathcal{PG}$ be two q-tame persistence modules. Then*

$$d_B(\mathcal{PF}, \mathcal{PG}) \leq d_I(\mathcal{PF}, \mathcal{PG}).$$

Corollary. *For two functions $f, g : \mathbb{X} \to \mathbb{R}$, if $\mathcal{P}(f)$ and $\mathcal{P}(g)$ are q-tame, then*

$$d_B(\mathcal{P}(f), \mathcal{P}(g)) \leq \|f - g\|_\infty .$$

Corollary ([2, Theorem 5.2]). *For two totally bounded metric spaces $\mathcal{X}, \mathcal{Y}$,*

$$d_I(\mathcal{PC}_{\mathcal{X}}(\mathcal{X}), \mathcal{PC}_{\mathcal{Y}}(\mathcal{Y})) \leq d_{GH}(\mathcal{X}, \mathcal{Y}).$$
$$d_I(\mathcal{PR}(\mathcal{X}), \mathcal{PR}(\mathcal{Y})) \leq d_{GH}(\mathcal{X}, \mathcal{Y}).$$

..

References

[1] Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. *The structure and stability of persistence modules.* SpringerBriefs in Mathematics. Springer, [Cham], 2016.

[2] Frederic Chazal, Vin de Silva, and Steve Oudot. Persistence stability for geometric complexes, 2013.

[3] Herbert Edelsbrunner and John L. Harer. *Computational topology.* American Mathematical Society, Providence, RI, 2010. An introduction.

[4] Allen Hatcher. *Algebraic topology.* Cambridge University Press, Cambridge, 2002.