INF 556: Topological Data Analysis **Fall 2018**

Lecture 5: Introduction to persistent homology

Lecturer: Steve Oudot *T.A.: Théo Lacombe*

Disclaimer:

Some typo and errors may remain. Please mention them at `theo.lacombe@polytechnique.edu`. Use these notes with caution, especially during the exam (we decline all responsibility linked with the use of these notes during the exam session).

Reminder: These notes are a concise summary of the lectures. They do not intend in any case to substitute to your personal notes and are just an additional support in order to clarify or insist on some points.

Keywords: filtration, persistence module, interval module, decomposition, persistence diagram, bottleneck distance, stability theorem.

5.1 Filtrations and persistence module

Let $T \subset \mathbb{R}$ be a set of indices.

Definition 1. *A* filtration *over T is a family $\mathcal{F} = (F_t)_{t \in T}$ of increasing (for inclusion) topological spaces:*

$$\forall t, t' \in T, t \leqslant t' \Rightarrow F_t \subset F_{t'}$$

Examples

- sublevel sets of a function $f : \mathbb{X} \to \mathbb{R}$: $F_t := f^{-1}((-\infty, t])$
- superlevel sets of $f : \mathbb{X} \to \mathbb{R}$: $F_t := f^{-1}([t, +\infty))$
- *offsets* (i.e. sublevel sets of the distance function) to a compact $K \in \mathbb{R}^d$:

  $$\forall t \in \mathbb{R}_+, F_t := \bigcup_{x \in K} B(x, t)$$

  where $B(x, t)$ is the closed euclidean ball centered in x with radius t.

Goal: Encode and estimate the evolution of the topology throughout scales of a family of topological spaces, for $t \in T$ in increasing order.

To do so, the idea is to use *homology* of F_t for each t, and we got a *persistence module.*

Definition 2 (Persistence module)**.** *Let $\mathbb{K}$ be a given field. A* persistence module *over $T \subset \mathbb{R}$ is a family $\mathbb{V} = (V_t)_{t \in T}$ of $\mathbb{K}$-vector spaces endowed with linear application $v_t^{t'} : V_t \to V_{t'}$ such that:*

$$\forall t \in T, v_t^t = id$$
$$\forall t \leqslant t' \leqslant t'' \in T, v_{t'}^{t''} \circ v_t^{t'} = v_t^{t''}$$

This condition are called functoriallity *conditions.*

$$\begin{tikzcd}
& V_{t'} \arrow[loop above, "\mathrm{id}_{V_{t'}}"] \arrow[dr, "v_{t'}^{t''}"] & \\
V_t \arrow[loop left, "\mathrm{id}_{V_t}"] \arrow[ur, "v_t^{t'}"] \arrow[rr, "v_t^{t''}"] & & V_{t''} \arrow[loop right, "\mathrm{id}_{V_{t''}}"]
\end{tikzcd}$$

Figure 5.1: Schema of functoriallity condition for persistence module: loop must be identity, and the oriented edges should "commute", that is the path you take to go from a vertex to another do not depends on the intermediate vertices you take.

Reminder: The homology verifies functorial properties: if you have: $X \xrightarrow{f} Y \xrightarrow{g} Z$, then $(f \circ g)_* = f_* \circ g_*$.

Link with filtrations: If we have $\mathcal{F} = (F_t)_t$ a filtration, we can apply the homology functor H_*:

- $\forall t \in T$, we define $V_t := H_*(F_t, \mathbb{K})$
- $\forall t \leqslant t'$, let $v_t^{t'}$ be the linear application induced by the canonical inclusion $F_t \overset{i}{\hookrightarrow} F_{t'}$.

One can check that this define a persistence module on our filtration.

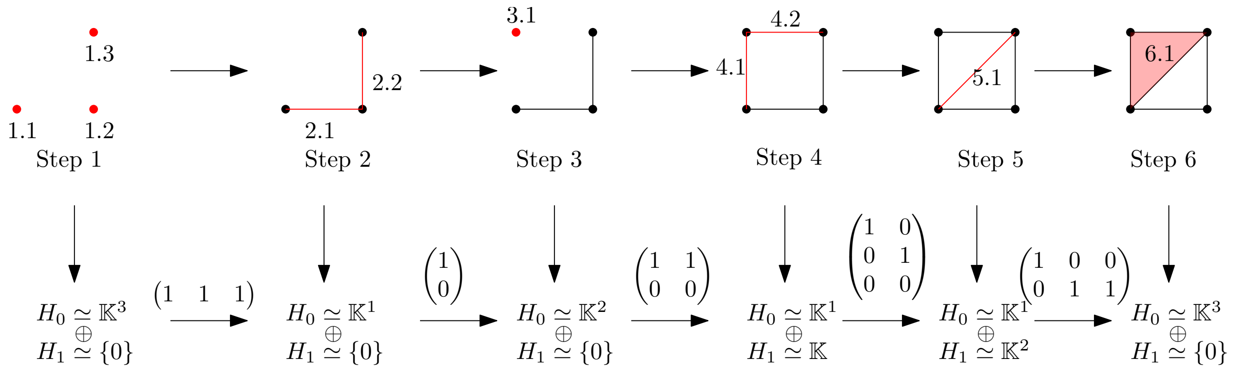

1.3
1.1
1.2
Step 1
2.2
2.1
Step 2
3.1
Step 3
4.2
4.1
Step 4
5.1
Step 5
6.1
Step 6
$H_0 \simeq \mathbb{K}^3 \oplus H_1 \simeq \{0\}$
$\begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$
$H_0 \simeq \mathbb{K}^1 \oplus H_1 \simeq \{0\}$
$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$
$H_0 \simeq \mathbb{K}^2 \oplus H_1 \simeq \{0\}$
$\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$
$H_0 \simeq \mathbb{K}^1 \oplus H_1 \simeq \mathbb{K}$
$\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$
$H_0 \simeq \mathbb{K}^1 \oplus H_1 \simeq \mathbb{K}^2$
$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}$
$H_0 \simeq \mathbb{K}^3 \oplus H_1 \simeq \{0\}$

Figure 5.2: An example of filtration over a simplicial complex, and the corresponding persistence module, that is $H_0(F_t) \oplus H_1(F_t), t \in \{1 \dots 6\}$. Linear applications $v_t^{t+1} : H_0(F_t) \oplus H_1(F_t) \to H_0(F_{t+1}) \oplus H_1(F_{t+1})$ are represented as matrices at each step.

Goal: Summarize / encode the algebraic structure of the persistence module $\mathbb{V}$ with a *barcode*.

5.2 Decompositions

$T, \mathbb{K}$ are given.

Our goal is to *decompose* $\mathbb{V}$, that is to be able to write $\mathbb{V} = \mathbb{V}_1 \oplus \dots \oplus \mathbb{V}_m$, where $\mathbb{V}_i$ are simple and cannot be decomposed.[1]

[1] You can think about it as some analogy with the decomposition of an integer into a product of prime numbers (which cannot - by definition - be decomposed again).

Definition 3. *An* interval *of T is a subset $I \subset T$ such as:*

$$\forall t \leqslant t' \leqslant t'' \in T : t, t'' \in I \Rightarrow t' \in I$$

Definition 4 (interval module)**.** *An interval module over $I \subset T$ is a persistence module $\mathbb{V}$ defined by:*

- *$V_t = \mathbb{K}$ if $t \in I$, $V_t = \{0\}$ otherwise.*
- *$\forall t, t' \in T, v_t^{t'} = \mathrm{id}_{\mathbb{K}}$ if $t, t' \in I$, $v_t^{t'} = 0$ otherwise.*

Notation for $I = [b, d]$:

$$\mathbb{I}_{[b,d]} := \underbrace{\{0\} \xrightarrow{0} \dots \xrightarrow{0} \{0\}}_{t<b} \xrightarrow{0} \underbrace{\mathbb{K} \xrightarrow{\mathrm{id}} \dots \xrightarrow{\mathrm{id}} \mathbb{K}}_{b \leqslant t \leqslant d} \xrightarrow{0} \underbrace{\{0\} \xrightarrow{0} \dots \xrightarrow{0} \{0\}}_{d<t}$$

Analogous definitions and notations stand for $I = (a, b], (a, b), [a, b)$, etc.

Idea: These interval module are "basic bricks" for decomposition of persistence module and, with some assumption, we will have decomposition theorems which will state that module can be decomposed into a direct sum (see below) of interval module.

Definition 5 (Direct sum of module)**.** *Given $\mathbb{V} = (V_t)_t, \mathbb{W} = (W_t)_t$ two persistence module with corresponding linear application $(v_t^{t'})_{t,t'}, (w_t^{t'})_{t,t'}$, we define $\mathbb{V} \oplus \mathbb{W}$ by:*

- $\mathbb{V} \oplus \mathbb{W} = (V_t \oplus W_t)_t$
- *The corresponding linear applications are denoted by:*

$$\begin{aligned} (v \oplus w)_t^{t'} &: V_t \oplus W_t \to V_{t'} \oplus W_{t'} \\ & (x, y) \mapsto (v_t^{t'}(x), w_t^{t'}(y)) \end{aligned}$$

This definition extends naturally for a family of persistence module $(\mathbb{V}_j)j \in J$, denoted by $\mathbb{V} := \bigoplus_{j \in J} \mathbb{V}_j$.

One can easily check that $\mathbb{V} \oplus \mathbb{W}$ is also a persistence module (idem for a family).

Theorem 1. *A persistence module $\mathbb{V}$ can be decomposed as a direct sum of interval module, written as:*

$$\mathbb{V} \simeq \bigoplus \mathbb{I}_{[b_j, d_j]}$$

in the following case (sufficient, not necessary):

1. *If T is finite. [Gabriel, 72].*
2. *When all the vector spaces V_t are finite-dimensional, [Crowley-Boevey, 2012].*

Furthermore, when it exists, the decomposition is unique (up to isomorphism and ordering of terms).

When we have such a decomposition, knowing the intervals $[b_j, d_j]$ gives a complete description of the structure of $\mathbb{V}$.

However, requiring these conditions may be too restrictive for our applications. One can consider

$$X = \{0\} \cup \bigcup_{n \geqslant 1} \left\{ \frac{1}{n} \right\} \subset \mathbb{R}$$

along with:

$$F_t := [-t, t] \cup \bigcup_{n \geqslant 1} \left[\frac{1}{n} - t, \frac{1}{n} + t \right]$$

$$V_t := H_0(F_t) \simeq \mathbb{K}^{n_t} \qquad \text{where } n_t = \left| \left\{ n \in \mathbb{N}^* \middle| \frac{1}{n+1} + t < \frac{1}{n} - t \right\} \right|$$

and observe that:

- V_0 is infinite-dimensional
- V_t is finite-dimensional (for all $t > 0$)
- $T = \mathbb{R}_+$

So none of the two previous conditions is satisfied, but $\mathbb{V}$ can still be decomposed as:

$$\mathbb{V} \simeq \mathbb{I}_{[0,+\infty)} \oplus \bigoplus_{n \geqslant 1} \mathbb{I}_{\left[0, \frac{1}{2n(n+1)}\right)}$$

Definition 6. *A persistence module $\mathbb{V}$ is said to be* q-tame *if $\forall t < t' \in T$, $\operatorname{rank}(v_t^{t'})$ is finite.*

Theorem 2 (Chazal, Cohen-Steiner, Glisse, Guibas, O., 2009)**.** *A persistence module $\mathbb{V} = (V_t)_t$ admits a well-defined barcode as soon as $\mathbb{V}$ is q-tame, even if $\mathbb{V}$ is not decomposable.*

Examples: As a good news, most of the filtrations we will consider will induce q-tame persistence module. It concerns (among other filtrations):

- Offsets (distances to a compact of $\mathbb{R}^d$)
- All Morse functions
- Sublevel and superlevel sets of functions $f : \mathbb{X} \to \mathbb{R}$ with $\mathbb{X}$ triangulable.

5.3 Computations of barcodes and persistence diagrams

Input: A simplicial filtration, that is a filtration over a simplicial complex K which verifies:

- $T = \{0, 1, \ldots, m\}$ (finite, so we have a decomposable filtration).
- $K_0 = \emptyset, K_m = K$
- $\forall t \in T, K_t$ is a simplicial complex, which is a sub-complex of K_{t+1}.

We can also assume (easy to verify) that we actually only add one simplex at each step, that is $K_{t+1} \backslash K_t = \{\sigma_t\}$.

Algorithm: We can basically apply the same algorithm as for simplicial homology, just by adding the ordering over the simplices:

1. Write the matrix M of the boundary operator ∂
2. define
$$\text{low}(j) := \begin{cases} \max\{i | M_{ij} \neq 0\} \\ 0 \text{ if } M_{ij} = 0 \text{ for all } i \end{cases}$$
3. Use Gaussian elimination from left to the right.

Algorithm 1 Compute the barcode corresponding to a simplicial filtration

```
for $j = 1 \ldots m$ do
  while $\exists i < j$ s.t. $\text{low}(i) = \text{low}(j) \neq 0$ do
    $c_j \leftarrow c_j - \frac{M[\text{low}(i), j]}{M[\text{low}(i), i]} c_i$
  end while
end for
return The reduced matrix.
```

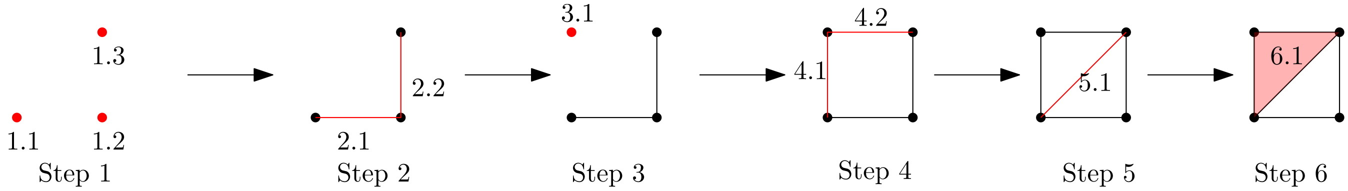
4. Interpretation: (after reduction)

   - Each column with full 0 entries induces a cycle.
   - Any other column j (with non 0) induces a boundary which trivialize (destroy) the cycle $i = \text{low}(j)$

   Thus:

   - Each column i with full 0 encodes the beginning of one interval module in the decomposition of the persistence module $H_*(K)$, i.e. $\partial\widehat{\sigma}_i \leftrightarrow$ module interval $\mathbb{I}_{[i,?]}$.
   - Finite intervals are $\mathbb{I}_{[i,j]}$ with $i = \text{low}(j)$
   - Infinite intervals are $\mathbb{I}_{[i,+\infty)}$ when there is no j such that $i = \text{low}(j)$.

Example: $\mathbb{K} = \mathbb{Z}/2\mathbb{Z}$, we take the complex introduced in figure 5.2



Leading to (matrix reduction):

$$\begin{array}{c|cccccc|}
 & 2.1 & 2.2 & 4.1 & 4.2 & 5.1 & 6.1 \\
1.1 & 1 & & & 1 & 1 & \\
1.2 & \boxed{1} & 1 & & & & \\
1.3 & & \boxed{1} & 1 & & 1 & \\
2.1 & & & & & & \\
2.2 & & & & & & \\
3.1 & & & \boxed{1} & \boxed{1} & & \\
4.1 & & & & & & 1 \\
4.2 & & & & & & 1 \\
5.1 & & & & & & \boxed{1} \\
6.1 & & & & & &
\end{array}
\xrightarrow{\text{Gaussian elimination}}
\begin{array}{c|cccccc|}
 & 2.1 & 2.2 & 4.1 & 4.2 & 5.1 & 6.1 \\
1.1 & 1 & & & & & \\
1.2 & \boxed{1} & 1 & & & & \\
1.3 & & \boxed{1} & 1 & & & \\
2.1 & & & & & & \\
2.2 & & & & & & \\
3.1 & & & \boxed{1} & & & \\
4.1 & & & & & & 1 \\
4.2 & & & & & & 1 \\
5.1 & & & & & & \boxed{1} \\
6.1 & & & & & &
\end{array}$$

Thus,

$$H_0 \simeq \mathbb{I}_{[1.1,+\infty)} \oplus \mathbb{I}_{[1.2,2.1)} \oplus \mathbb{I}_{[1.3,2.2)} \oplus \mathbb{I}_{[3.1,4.1)}$$
$$H_1 \simeq \mathbb{I}_{[4.2,+\infty)} \oplus \mathbb{I}_{[5.1,6.1)}$$

i.e., if we filter just for $T = \{1, 2, \dots, 6\}$ for the steps:

$$H_0 \simeq \mathbb{I}_{[1,+\infty)} \oplus \mathbb{I}_{[1,2)} \oplus \mathbb{I}_{[1,2)} \oplus \mathbb{I}_{[3,4)}$$
$$H_1 \simeq \mathbb{I}_{[4,+\infty)} \oplus \mathbb{I}_{[5,6)}$$

5.4 Stability

Persistence diagrams: The goal of this section is to define a way to compare barcode in order to give some powerful results. Due to the distances involved (see below), it is more convenient to deal with an equivalent representation of the barcode, which is called the *persistence diagram* of our persistence module.

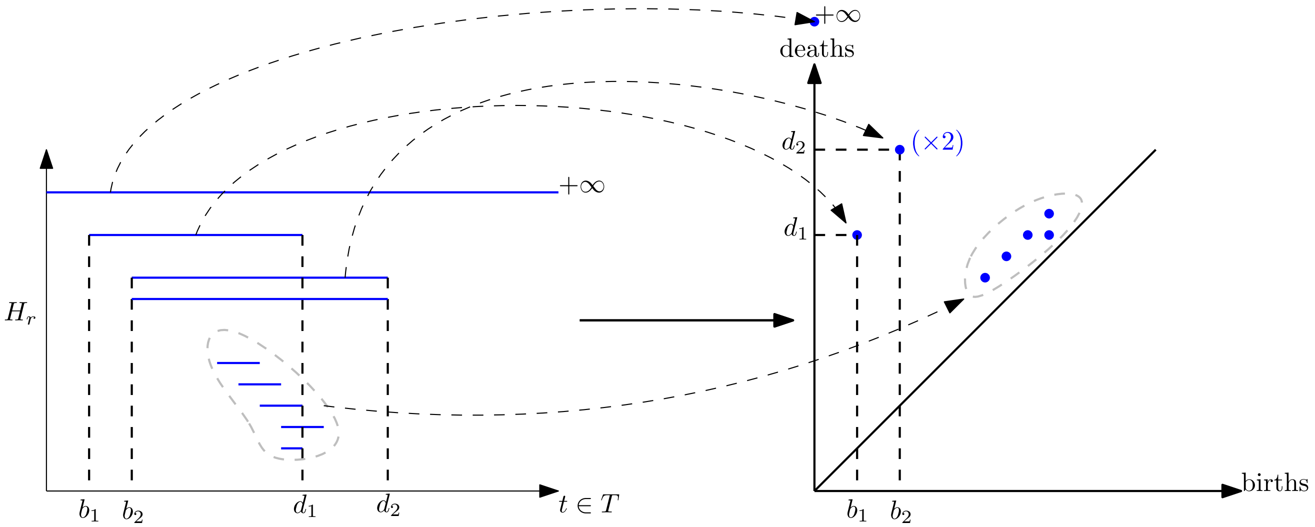


Figure 5.3: *(left)* Barcode. *(right)* The corresponding persistence diagram. To each interval $[b_j, d_j]$ in the barcode corresponds one point of coordinate (b_j, d_j) in the persistence diagram. Note that there could be several points with same coordinates (b, d). In this case, points are counted with multiplicity. Thus, a diagram is often referred to as a *multiset* of points above the diagonal.

Goal: We have an operator $f \mapsto \mathrm{Dg}(f)$ which maps a function to a diagram (for example by taking the sublevel sets). The question is to know if this operator has some stability properties: if we modify f, does $\mathrm{Dg}(f)$ changes a lot?

In order to quantify that, we need metrics on input and output spaces.

- For functions $f, g : \mathbb{X} \to \mathbb{R}$, we compare them with $\|.\|_\infty$
- For diagrams, we need to introduce a matching distance which is called the *bottleneck distance.*

Definition 7 (Partial matching)**.** *A partial matching between two set A, B is a subset $M \subset A \times B$ so that:*

- *$\forall a \in A$, there is at most one b such that $(a, b) \in M$*
- *$\forall b \in B$, there is at most one a such that $(a, b) \in M$*

Definition 8. *We define $\Delta = \{(x, x) \in \mathbb{R}^2\}$ (the diagonal). We note $\Omega = \{(x, y) \in \mathbb{R}^2 | x < y\}$ (the points above the diagonal). We define the following cost function on $\overline{\Omega}$ (here, the diagonal Δ must be understood as **one** additional point):*

- *$c(x, y) = \|x - y\|_\infty$ for $x, y \in \Omega$.*
- *$c(x, \Delta) = c(\Delta, x) = \|x - \pi_\Delta(x)\|_\infty$, where $\pi_\Delta(x)$ is the orthogonal projection of x onto the diagonal Δ.*
- *$c(\Delta, \Delta) = 0$*

Definition 9 (Matching cost)**.** *Let A, B be two sets of points in Ω (i.e. diagrams) and M a partial matching between A and B. Its cost is defined as:*

$$c(M) = \max \left\{ \sup_{(a,b) \in M} c(a, b), \sup_{(s) \in A \cup B \ unmatched} c(s, \Delta) \right\}$$

Definition 10 (Bottleneck distance)**.** *Let A, B be two diagrams. The bottleneck distance between these two diagrams is defined as:*

$$d_B^\infty := \inf_{M : A \leftrightarrow B} c(M) \tag{5.1}$$

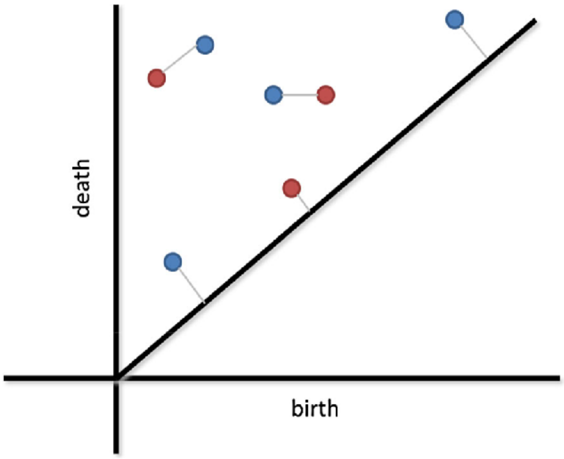


Figure 5.4: An example of (optimal) matching between two diagrams (blue and red dots). The bottleneck distance between these two diagrams is, by definition, the length of the longest edge (measured in $\|.\|_\infty$).

Why the bottleneck? The following theorem states that the bottleneck is a *good* metric to manipulate diagrams due to a stability result.

Theorem 3 (Cohen-Steiner, Edelsbrunner, Harer 2005, Chazal, Cohen-Steiner, Glisse, Guibas, O., 2009)**.** *Let $f, g : \mathbb{X} \to \mathbb{R}$ be two* q-tame *functions (i.e. the homology module they induce are q-tame), and $Dg(f), Dg(g)$ be the persistence diagrams they induce. The operator $f \mapsto Dg(f)$ is* 1*-Lipschitz, that is:*

$$d_B^\infty(Dg(f), Dg(g)) \leqslant \|f - g\|_\infty \tag{5.2}$$

Proof. Cf slides page 6. Cf Book, chapter 3, section 2.1 and 2.2.

It basically use the following interleaving:

$$\begin{tikzcd}
{} \arrow[r, "\subseteq"] & F_{(2n-2)\epsilon} \arrow[rr, "\subseteq"] \arrow[dr, "\subseteq"] & & F_{2n\epsilon} \arrow[rr, "\subseteq"] \arrow[dr, "\subseteq"] & & F_{(2n-2)\epsilon} \arrow[r, "\subseteq"] & {} \\
& {} \arrow[r, "\subseteq"] & G_{(2n-1)\epsilon} \arrow[rr, "\subseteq"] \arrow[ur, "\subseteq"] & & G_{(2n+1)\epsilon} \arrow[r, "\subseteq"] \arrow[ur, "\subseteq"] & {} &
\end{tikzcd}$$

In homology, it becomes:

$$\begin{tikzcd}
{} \arrow[r] & H_*(F_{(2n-2)\epsilon}) \arrow[rr] \arrow[dr] & {} & H_*(F_{2n\epsilon}) \arrow[rr] \arrow[dr] \arrow[d, phantom, "\circlearrowleft"] & {} & H_*(F_{(2n-2)\epsilon}) \arrow[r] & {} \\
& {} \arrow[r] & H_*(G_{(2n-1)\epsilon}) \arrow[rr] \arrow[ur] \arrow[u, phantom, "\circlearrowleft"] & {} & H_*(G_{(2n+1)\epsilon}) \arrow[r] \arrow[ur] \arrow[u, phantom, "\circlearrowleft"] & {} &
\end{tikzcd}$$

□

Definition 11. *Let $\mathbb{V} = (V_t)_{t\in\mathbb{R}}$ and $\mathbb{W} = (W_t)_{t\in\mathbb{R}}$ be two persistence module on $\mathbb{R}$. $\mathbb{V}$ and $\mathbb{W}$ are said to be* interleaved *if there are two families of applications $\varphi = (\varphi_t)_t, \psi = (\psi_t)_t$ where $\varphi_t : V_t \to W_{t+\varepsilon}$ and $\psi_t : W_t \to V_{t+\varepsilon}$ such that for all $t \leqslant t'$, the following diagrams commute:*

$$\begin{tikzcd}
V_t \arrow[r, "v_t^{t'}"] \arrow[dr, "\varphi_t"'] & V_{t'} \arrow[dr, "\varphi_{t'}"] & \\
& W_t \arrow[r, "w_{t+\epsilon}^{t'+\epsilon}"] & W_{t'+\epsilon}
\end{tikzcd}
\qquad
\begin{tikzcd}
& V_{t+\epsilon} \arrow[r] & V_{t'+\epsilon} \\
W_t \arrow[ur, "\psi_t"] \arrow[r] & W_{t'} \arrow[ur, "\psi_{t'}"'] &
\end{tikzcd}$$

$$\begin{tikzcd}
V_t \arrow[rr] \arrow[dr, "\varphi_t"'] & & V_{t+2\epsilon} \\
& W_{t+\epsilon} \arrow[ur, "\psi_{t+\epsilon}"'] &
\end{tikzcd}
\qquad
\begin{tikzcd}
& V_{t+\epsilon} \arrow[dr, "\varphi_{t+\epsilon}"] & \\
W_t \arrow[ur, "\psi_t"] \arrow[rr] & & W_{t+2\epsilon}
\end{tikzcd}$$

Figure 5.5: Schematic representation of interleaving between persistence module.

Theorem 4 (Chazal, Cohen-Steiner, Glisse, Guibas, O., 2009)**.** *If* $\mathbb{V}, \mathbb{W}$ *are q-tame and* ε*-interleaved, then:*

$$d_B^{\infty}(Dg(\mathbb{V}), Dg(\mathbb{W})) \leqslant \varepsilon$$

Corollary 1. *let* $d_i(\mathbb{V}, \mathbb{W}) := \inf\{\varepsilon \geqslant 0 | \mathbb{V}, \mathbb{W}$ *are interleaved* $\} \in \mathbb{R}_+ \cup \{+\infty\}$.

Then, for all q-tame module $\mathbb{V}, \mathbb{W}$,

$$d_B(Dg(\mathbb{V}), Dg(\mathbb{W})) \leqslant d_i(\mathbb{V}, \mathbb{W})$$

Actually, we even have this isometric property [Lesnick 2011]:

$$d_B(Dg(\mathbb{V}), Dg(\mathbb{W})) = d_i(\mathbb{V}, \mathbb{W})$$

Conclusion: Persistence diagrams, endowed with the bottleneck distance, are stable representations of the underlying topology (persistent homology) of our data, and this is a natural metric regarding the standard one between module.