PERSISTENT HOMOLOGY

Contents

Persistent homology is a powerful tool to compute, study and encode efficiently multiscale topological features of nested families of simplicial complexes and topological spaces. It does not only provide efficient algorithms to compute the Betti numbers of each complex in the considered families, as required for homology inference in the previous section, but also encodes the evolution of the homology groups of the nested complexes across the scales.

1. Filtrations

Definition 1.1 (Filtration)**.** A *filtration of a simplicial complex* $\mathcal{K}$ is a nested family of subcomplexes $(\mathcal{K}_r)_{r\in T}$, where $T \subset \mathbb{R}$, such that for any $r, r' \in T$, if $r \leqslant r'$ then $\mathcal{K}_r \subset \mathcal{K}_{r'}$, and $\mathcal{K} = \cup_{r\in T}\mathcal{K}_r$.

More generally, a *filtration* of a topological space M is a nested family of subspaces $(M_r)_{r\in T}$, where $T \subset \mathbb{R}$, such that for any $r, r' \in T$, if $r \leqslant r'$ then $M_r \subset M_{r'}$ and, $M = \cup_{r\in T} M_r$. For example, if $f : M \to \mathbb{R}$ is a function, then the family $M_r = f^{-1}((-\infty, r])$, $r \in \mathbb{R}$ defines a filtration called the sublevel set filtration of f.

Remark 1.2. (i) The subset T may be either finite or infinite.
(ii) In practical situations, the parameter $r \in T$ can often be interpreted as a scale parameter and filtrations classically used in TDA often belong to one of the two following families.

Example 1.3 (Filtrations Built on Top of Data)**.** Given a subset $\mathbb{X}$ of a compact metric space (M, ρ), the families of Rips-Vietoris complexes $(\mathrm{Rips}(\mathbb{X}, r))_{r\in\mathbb{R}}$ and and Čech complexes $(\check{\mathrm{C}}\mathrm{ech}(\mathbb{X}, r))_{r\in\mathbb{R}}$ are filtrations[1]. Here, the parameter r can be interpreted as a resolution at which one considers the data set $\mathbb{X}$.

In particular, if $\mathbb{X}$ is a point cloud in $\mathbb{R}^d$, thanks to the Nerve theorem, the filtration $(\check{\mathrm{C}}\mathrm{ech}(\mathbb{X}, r))_{r\in\mathbb{R}}$ encodes the topology of the whole family of unions of balls $\mathbb{X}^r = \cup_{x\in\mathbb{X}}\mathrm{B}(x, r)$, as r goes from 0 to $+\infty$.

Example 1.4 (Sublevel Sets Filtrations)**.** Functions defined on the vertices of a simplicial complex give rise to another important example of filtration: let $\mathcal{K}$ be a simplicial complex with vertex set V and $f : V \to \mathbb{R}$. Then f can be extended to all simplices of $\mathcal{K}$ by $f([v_0, \cdots, v_k]) = \max_{1\leqslant i\leqslant k} f(v_i)$ for any simplex $\sigma = [v_0, \cdots, v_k] \in \mathcal{K}$. The family of subcomplexes $\mathcal{K}_r = \{\sigma \in \mathcal{K} | f(\sigma) \leqslant r\}$ defines a filtration call the sublevel set filtration of f. Similarly, one can define the upperlevel set filtration of f.

In practice, even if the index set is infinite, all the considered filtrations are built on finite sets and are indeed finite. For example, when $\mathbb{X}$ is finite, the Vietoris-Rips complex $\mathrm{Rips}(\mathbb{X}, r)$ changes only at a finite number of indices r. This allows to easily handle them from an algorithmic perspective.

2. Starting with a Few Examples

Given a filtration $\mathrm{Filt} = (F_r)_{r\in T}$ of a simplicial complex or a topological space, the homology of F_r changes as r increases: new connected components can appear, existing component can merge, loops and cavities can appear or be filled, etc. Persistent homology tracks these changes, identifies the appearing features and associates a life time to them. The resulting information is encoded as a set of intervals called a *barcode* or, equivalently, as a multiset of points in $\mathbb{R}^2$ where the coordinate of each point is the starting and end point of the corresponding interval.

Before giving formal definitions, we introduce and illustrate persistent homology on three simple examples.

Example 2.1 (Smooth Real Function)**.** Let $f : [0, 1] \to \mathbb{R}$ be the function of Figure 1 and let $(F_r = f^{-1}((-\infty, r]))_{r\in\mathbb{R}}$ be the sublevel set filtration of f.

(a_1) All the sublevel sets of f are either empty or a union of interval, so the only non trivial topological information they carry is their 0-dimensional homology, i.e. their number of connected components. For $r < a_1$, F_r is empty, but at $r = a_1$ a first connected component appears in F_{a_1}. Persistent homology thus registers a_1 as the birth time of a connected component and start to keep track of it by creating an interval starting at a_1.

(a_2) Then, F_r remains connected until r reaches the value a_2 where a second connected component appears. Persistent homology starts to keep track of this new connected component by creating a second interval starting at a_2.

[1] we take here the convention that for $r < 0$, $\mathrm{Rips}(\mathbb{X}, r) = \check{\mathrm{C}}\mathrm{ech}(\mathbb{X}, r) = \emptyset$

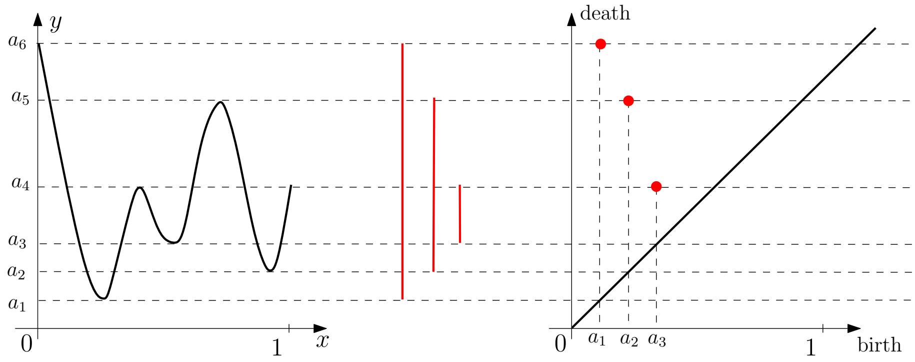


FIGURE 1. The persistence barcode and the persistence diagram of a function $f : [0,1] \to \mathbb{R}$.

(a_3) Similarly, when r reaches a_3, a new connected component appears and persistent homology creates a new interval starting at a_3.

(a_4) When r reaches a_4, the two connected components created at a_1 and a_3 merges together to give a single larger component. At this step, persistent homology follows the rule that this is the most recently appeared component in the filtration that dies: the interval started at a_3 is thus ended at a_4 and a first persistence interval encoding the lifespan of the component born at a_3 is created.

(a_5) When r reaches a_5, as in the previous case, the component born at a_2 dies and the persistent interval (a_2, a_5) is created.

(a_6) The interval created at a_1 remains until the end of the filtration giving rise to the persistent interval (a_1, a_6) if the filtration is stopped at a_6, or $(a_1, +\infty)$ if r goes to $+\infty$ (notice that in this later case, the filtration remains constant for $r > a_6$).

The obtained set of intervals encoding the span life of the different homological features encountered along the filtration is called the *persistence barcode* of f. Each interval (a, a') can be represented by the point of coordinates (a, a') in $\mathbb{R}^2$ plane. The resulting set of points is called the *persistence diagram* of f. Notice that a function may have several copies of the same interval in its persistence barcode. As a consequence, the persistence diagram of f is indeed a multi-set where each point has an integer valued multiplicity. Last, for technical reasons that will become clear in the next section, one adds to the persistence all the points of the diagonal $\Delta = \{(b,d) : b = d\}$ with an infinite multiplicity.

Example 2.2 (Surface in Space)**.** Let now $f : M \to \mathbb{R}$ be the function of Figure 2 where M is a 2-dimensional surface homeomorphic to a torus, and let $(F_r = f^{-1}((-\infty, r]))_{r\in\mathbb{R}}$ be the sublevel set filtration of f. The 0-dimensional persistent homology is computed as in the previous example, giving rise to the red bars in the barcode. Now, the sublevel sets also carry 1-dimensional homological features.

(a_1) When r goes through the height a_1, the sublevel sets F_r that were homeomorphic to two discs become homeomorphic to the disjoint union of a disc and an annulus, creating a first cycle homologous to σ_1 on Figure

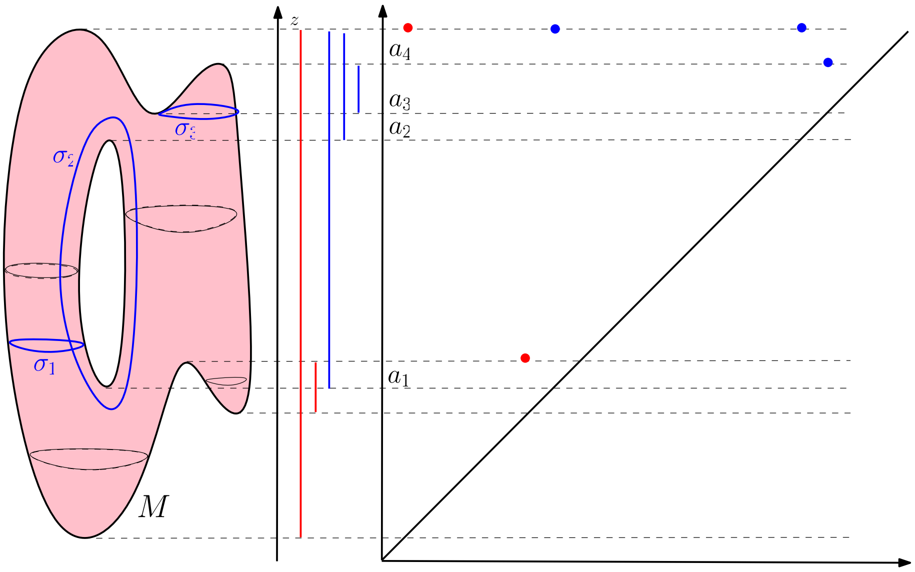


Figure 2. The persistence barcode and the persistence diagram of the height function (projection on the z-axis) defined on a surface in $\mathbb{R}^3$.

2. A interval (in blue) representing the birth of this new 1-cycle is thus started at a_1.

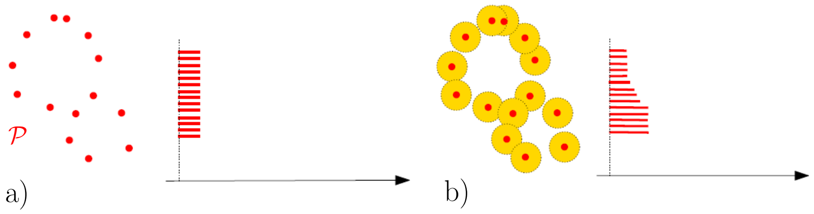
(a_2) Similarly, when r goes through the height a_2 a second cycle, homologous to σ_2 is created, giving rise to the start of a new persistent interval. These two created cycles are never filled (indeed they span $H_1(M)$) and the corresponding intervals remains until the end of the filtration.

(a_3) When r reaches a_3, a new cycle σ_3 is created.

(a_4) This cycle is filled and thus dies at a_4, giving rise to the persistence interval (a_3, a_4).

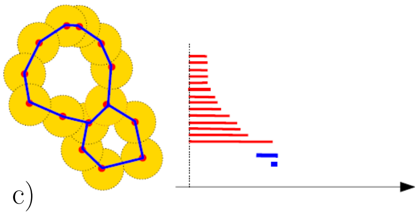
So, now, the sublevel set filtration of f gives rise to two barcodes, one for 0-dimensional homology (in red) and one for 1-dimensional homology (in blue). As previously, these two barcodes can equivalently be represented as diagrams in the plane.

Example 2.3 (Offsets of a Point Cloud)**.** In this last example we consider the filtration given by a union of growing balls centered on the finite set of points $\mathcal{P}$, as pictured below. Notice that this is the sublevel set filtration of the distance function to $\mathcal{P}$, that is $(F_r = \mathrm{d}_{\mathcal{P}}^{-1}((-\infty, r]))_{r\in\mathbb{R}}$. Thanks to the Nerve Theorem, this filtration is homotopy equivalent to the Čech filtration built on top of $\mathcal{P}$.

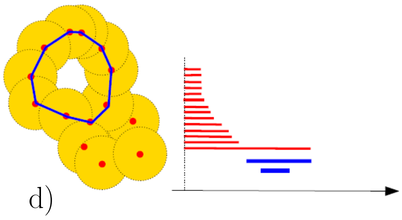


a) For the radius $r = 0$, the union of balls is reduced to the initial finite set of point, each of them corresponding to a 0-dimensional feature, i.e. a connected component; an interval is created for the *birth* for each of these features at $r = 0$.

b) Some of the balls started to overlap resulting in the *death* of some connected components that get merged together; the persistence diagram keeps track of these deaths, putting an end point to the corresponding intervals as they disappear.



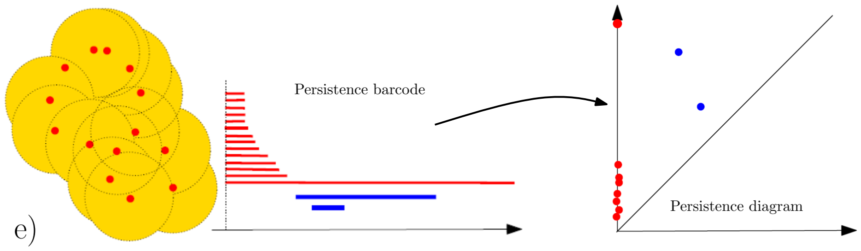
c)



d)

c) New components have merged giving rise to a single connected component and, so, all the intervals associated to a 0-dimensional feature have been ended, except the one corresponding to the remaining components; two new 1-dimensional features, have appeared resulting in two new intervals (in blue) starting at their birth scale.

d) One of the two 1-dimensional cycles has been filled, resulting in its death in the filtration and the end of the corresponding blue interval.



e)

e) all the 1-dimensional features have died, it only remains the long (and never dying) red interval. As in the previous examples, the final barcode can also be equivalently represented as a persistence diagram where every interval (a, b) is represented by the the point of coordinate (a, b) in $\mathbb{R}^2$. Intuitively the longer is an interval in the barcode or, equivalently the farther from the diagonal is the corresponding point in the diagram, the more persistent, and thus relevant, is the corresponding homological feature across the filtration. Notice also that for a given radius r, the k-th Betti number of the corresponding union of balls is equal of the number of persistence intervals corresponding to k-dimensional homological features and containing r. So, the persistence diagram can be seen as a multiscale topological signature encoding the homology of the union of balls for all radii as well as its evolution across the values of r.

3. Persistent Modules and Persistence Diagrams

Persistent diagrams can be formally and rigorously defined in a purely algebraic way. This requires some care and we only give here the basic necessary notions, leaving aside technical subtleties and difficulties.

Definition 3.1 (Persistence Module)**.** A *persistence module* $\mathbb{V}$ over a subset $T \subset \mathbb{R}$ of the real numbers is an indexed family of vector spaces $(V_r \mid r \in T)$ and a doubly-indexed family of linear maps $(v_s^r : V_r \to V_s \mid r \leqslant s)$ which satisfy the composition law $v_t^s \circ v_s^r = v_t^r$ whenever $r \leqslant s \leqslant t$, and where v_r^r is the identity map on V_r.

Example 3.2. Let $\mathrm{Filt} = (F_r)_{r \in T}$ be a filtration of a simplicial complex or a topological space. Given an integer $k \geqslant 0$ and considering the homology groups $H_k(F_r)$ we obtain a family of vector spaces, and the inclusions $i_s^r : F_r \hookrightarrow F_s$,for $r \leqslant s$, induce linear maps $(i_s^r)_* : H_k(F_r) \to H_k(F_s)$ at the homology level. Furthermore, these maps satisfy $(i_t^r)_* = (i_t^s \circ i_s^r)_* = (i_t^s)_* \circ (i_s^r)_*$ for all $r \leqslant s \leqslant t$.

In many cases, a persistence module can be decomposed into a direct sum of *intervals* modules $\mathbb{I}_{(b,d)}$ of the form

$$\cdots \to 0 \to \cdots \to 0 \to \mathbb{Z}_2 \to \cdots \to \mathbb{Z}_2 \to 0 \to \cdots$$

where the maps $\mathbb{Z}_2 \to \mathbb{Z}_2$ are identity maps while all the other maps are 0. Denoting b (resp. d) the infimum (resp. supremum) of the interval of indices corresponding to non zero vector spaces, such a module can be interpreted as a feature that appears in the filtration at index b and disappear at index d. When a persistence module $\mathbb{V}$ can be decomposed as a direct sum of interval modules, one can show that this decomposition is unique up to reordering the intervals (see [CdSGO16, Theorem 2.7]). As a consequence, the set of resulting intervals is independent of the decomposition of $\mathbb{V}$ and is called the *persistence barcode* of $\mathbb{V}$.

Remark 3.3. As in the examples of the previous section, each interval (b, d) in the barcode can be represented as the point of coordinates (b, d) in the plane $\mathbb{R}^2$. The disjoint union of these points, together with the diagonal $\Delta = \{b = d\}$ is a multi-set called the *persistence diagram* of $\mathbb{V}$.

The following result gives sufficient conditions for a persistence module to be decomposable as a direct sum of interval modules.

Theorem 3.4. *Let $\mathbb{V}$ be a persistence module indexed by $T \subset \mathbb{R}$. If T is a finite set or if all the vector spaces V_r are finite-dimensional, then $\mathbb{V}$ is decomposable as a direct sum of interval modules.*

Proof. See [CdSGO16, Theorem 2.8]. □

As both conditions above are satisfied for the persistent homology of filtrations of finite simplicial complexes, an immediate consequence of this result is that the persistence diagrams of such filtrations are always well-defined.

Unfortunately, Theorem 3.4 is not sufficient for our purposes of general data analysis. Indeed, there exist compact sets whose offsets do not induce pointwise finite-dimensional persistence modules, such as $\mathbb{X} = \{0\} \cup_{n \geqslant 1}$

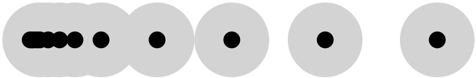


FIGURE 3. The set $\mathbb{X} = \{0\} \cup_{n\geqslant 1} \{1/n\}$ is compact, but $\beta_0(\mathbb{X}) = \infty$ and its offsets $(\mathbb{X}^r)_{r\geqslant 0}$ are naturally indexed by the infinite set $T = \mathbb{R}$.

$\{1/n\}$ (see Figure 3). However, it is possible to show that persistence diagrams can be defined as soon as the following simple condition is satisfied.

Definition 3.5 (q-tameness)**.** A persistence module $\mathbb{V}$ indexed by $T \subset \mathbb{R}$ is *q-tame* if for any $r < s$ in T, the rank of the linear map $v_s^r : V_r \to V_s$ is finite.

THEOREM 3.6 ([CdSGO16]). *If* $\mathbb{V}$ *is a q-tame persistence module, then it has a well-defined persistence diagram.*

Remark 3.7. (i) Theorem 3.6 is pretty strong, since its shows that the diagram is well-defined, even though $\mathbb{V}$ may not be interval-decomposable.

(ii) Such a persistence diagram $\mathrm{dgm}(\mathbb{V})$ is the union of the points of the diagonal Δ of $\mathbb{R}^2$, counted with infinite multiplicity, and a multi-set above the diagonal in $\mathbb{R}^2$ that is locally finite. Here, by locally finite we mean that for any rectangle R with sides parallel to the coordinate axes that does not intersect Δ, the number of points of $\mathrm{dgm}(\mathbb{V})$, counted with multiplicity, contained in R is finite.

(iii) (Insights on q-tameness) One can check [CdSGO16, Corollary 2.2] that the number of points in any rectangle $[a, b] \times [c, d]$ above the diagonal ($a \leqslant b \leqslant c \leqslant d$) corresponds to $\mathrm{rank}(v_b^c) - \mathrm{rank}(v_b^d) + \mathrm{rank}(v_a^d) - \mathrm{rank}(v_a^c)$. Letting $a \to -\infty$ and $d \to \infty$, we get that the number of points in the *quadrant* $(-\infty, b] \times [c, \infty)$ is finite whenever $c > b$, explaining the term *q-tame*.

The construction of persistence diagrams of q-tame modules is beyond the scope of this lesson but it gives rise to the same notion as in the case of decomposable modules. It can be done either by following the algebraic approach based upon the decomposability properties of modules, or by adopting a measure theoretic approach that allows to define diagrams as integer valued measures on a space of rectangles in the plane. We refer the reader to [CdSGO16] for more information. Although persistence modules encountered in practice are decomposable, the general framework of q-tame persistence module plays a fundamental role in the mathematical and statistical analysis of persistent homology.

Remark 3.8 (Diagram of a Filtration)**.** A filtration $\mathrm{Filt} = (F_r)_{r\in T}$ of a simplicial complex or of a topological space is said to be tame if for any integer k, the persistence module $(H_k(F_r) \mid r \in T)$ is q-tame. Notice that the filtrations of finite simplicial complexes are always tame. As a consequence, for any integer k a persistence diagram denoted $\mathrm{dgm}_k(\mathrm{Filt})$ is associated to the filtration Filt. When k is not explicitly specified and when there is no ambiguity, it is usual to drop the index k in the notation and to talk about "the" persistence diagram dgm(Filt) of the filtration Filt. This notation has to be understood as "$\mathrm{dgm}_k(\mathrm{Filt})$ for some k".

4. Metrics on the Space of Persistence Diagrams

To exploit the topological information and topological features inferred from persistent homology, one needs to be able to compare persistence diagrams, i.e. to endow the space of persistence diagrams with a metric structure. Although several metrics can be considered, the most fundamental one is known as the *bottleneck distance.*

Recall that a persistence diagram is the union of a discrete multi-set in the half-plane above the diagonal Δ and, for technical reasons that will become clear below, of Δ where the point of Δ are counted with infinite multiplicity.

Definition 4.1 (Matching)**.** A *matching* between two diagrams dgm_1 and dgm_2 is a subset $m \subset \mathrm{dgm}_1 \times \mathrm{dgm}_2$ such that every points in $\mathrm{dgm}_1 \setminus \Delta$ and $\mathrm{dgm}_2 \setminus \Delta$ appears exactly once in m.

In other words, for any $p \in \mathrm{dgm}_1 \setminus \Delta$, and for any $q \in \mathrm{dgm}_2 \setminus \Delta$, $(\{p\} \times \mathrm{dgm}_2) \cap m$ and $(\mathrm{dgm}_1 \times \{q\}) \cap m$ each contains a single pair, see Figure 4.

Definition 4.2 (Bottleneck Distance)**.** The *bottleneck distance* between dgm_1 and dgm_2 is then defined by

$$\mathrm{d_b}(\mathrm{dgm}_1, \mathrm{dgm}_2) = \inf_{\text{matching } m} \max_{(p,q)\in m} \|p - q\|_\infty .$$

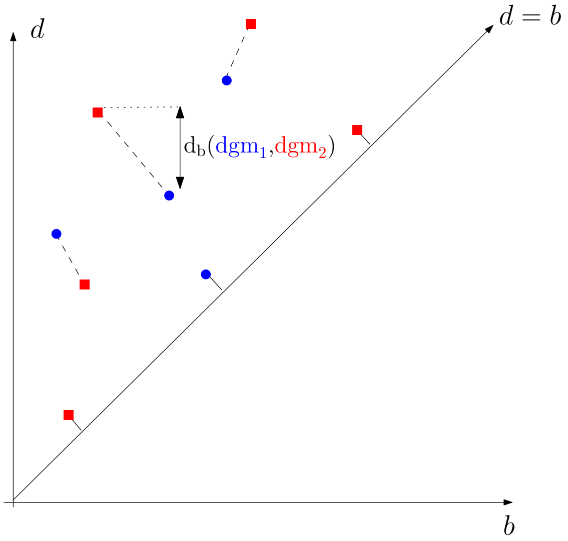


Figure 4. A perfect matching and the bottleneck distance between a blue and a red diagram. Notice that some points of both diagrams are matched to points of the diagonal.

Remark 4.3. (i) The practical computation of the bottleneck distance boils down to the computation of a perfect matching in a bipartite graph for which classical algorithms can be used.

(ii) The bottleneck metric is a L_∞-like metric. It turns out to be the natural one to express stability properties of persistence diagrams presented in Section 5, but it suffers from the same drawbacks as the usual L_∞ norms, i.e. it is completely determined by the largest distance among the pairs and do not take into account the closeness of the remaining pairs of points. A variant, to overcome this issue, the so-called Wasserstein distance between diagrams is sometimes considered. Given $p \geqslant 1$, it is defined by

$$\mathrm{W}_p(\mathrm{dgm}_1, \mathrm{dgm}_2)^p = \inf_{\text{matching } m} \sum_{(p,q)\in m} \|p-q\|_\infty^p.$$

Useful stability results for persistence in the metric W_p exist among the literature, but they rely on assumptions that make them consequences of the stability results in the bottleneck metric.

5. Stability

5.1. **A General Result.** A fundamental property of persistence homology is that persistence diagrams of filtrations built on top of data sets turn out to be very stable with respect to some perturbations of the data. To formalize and quantify such stability properties, we first need to precise the notion of perturbation that are allowed.

Rather than working directly with filtrations built on top of data sets, it turns out to be more convenient to define a notion of proximity between persistence module from which we will derive a general stability result for persistent homology. Then, most of the stability results for specific filtrations will appear as a consequence of this general theorem. To avoid technical discussions, from now on we assume, without loss of generality, that the considered persistence modules are indexed by $\mathbb{R}$.

Definition 5.1 (Homomorphism of Persistence Modules)**.** Let $\mathbb{V}, \mathbb{W}$ be two persistence modules indexed by $\mathbb{R}$. Given $\delta \in \mathbb{R}$, a *homomorphism of degree* δ between $\mathbb{V}$ and $\mathbb{W}$ is a collection Φ of linear maps $\phi_r : V_r \to W_{r+\delta}$, for all $r \in \mathbb{R}$ such that the following diagram commutes:

$$\begin{tikzcd}
V_r \arrow[rr, "v_s^r"] \arrow[dr, "\phi_r"] & & V_s \arrow[dr, "\phi_s"] & \\
& W_{r+\delta} \arrow[rr, "w_{s+\delta}^{r+\delta}"'] & & W_{s+\delta}
\end{tikzcd}$$

That is, for all $r \leqslant s$, $\phi_s \circ v_s^r = w_{s+\delta}^{r+\delta} \circ \phi_r$.

An important example of homomorphism of degree δ is the *shift endomorphism* $1_{\mathbb{V}}^{\delta}$ which consists of the families of linear maps $\phi_r = v_{r+\delta}^r$. Notice also that homomorphisms of persistence modules can naturally be composed: the composition of a homomorphism Ψ of degree δ between $\mathbb{U}$ and $\mathbb{V}$ and a homomorphism Φ of degree δ' between $\mathbb{V}$ and $\mathbb{W}$ naturally gives rise to a homomorphism $\Phi\Psi$ of degree $\delta + \delta'$ between $\mathbb{U}$ and $\mathbb{W}$.

Definition 5.2. Let $\delta \geqslant 0$. Two persistence modules $\mathbb{V}, \mathbb{W}$ are *δ-interleaved* if there exists two homomorphism of degree δ, Φ, from $\mathbb{V}$ to $\mathbb{W}$ and Ψ, from $\mathbb{W}$ to $\mathbb{V}$ such that $\Psi\Phi = 1_{\mathbb{V}}^{2\delta}$ and $\Phi\Psi = 1_{\mathbb{W}}^{2\delta}$.

$$\begin{tikzcd}
V_{r-\delta} \arrow[r, "v_r^{r-\delta}"'] \arrow[rr, bend left, "v_{r+\delta}^{r-\delta}"] \arrow[dr, "\phi_{r-\delta}"'] & V_r \arrow[r, "v_{r+\delta}^r"'] \arrow[dr, "\phi_r"] & V_{r+\delta} \\
W_{r-\delta} \arrow[r, "w_r^{r-\delta}"] \arrow[rr, bend right, "w_{r+\delta}^{r-\delta}"'] \arrow[ur, "\psi_{r-\delta}"] & W_r \arrow[r, "w_{r+\delta}^r"] \arrow[ur, "\psi_r"'] & W_{r+\delta}
\end{tikzcd}$$

Although it does not define a metric on the space of persistence modules, the notion of closeness between two persistence modules may be defined as the smallest non negative δ such that they are δ-interleaved. Moreover, it allows to formalize the following fundamental result.

Theorem 5.3 (Stability of Persistence). *Let $\mathbb{V}$ and $\mathbb{W}$ be two q-tame persistence modules. If $\mathbb{V}$ and $\mathbb{W}$ are δ-interleaved for some $\delta \geqslant 0$, then*

$$\mathrm{d_b}(\mathrm{dgm}(\mathbb{V}), \mathrm{dgm}(\mathbb{W})) \leqslant \delta.$$

Proof. See [CdSGO16]. □

Remark 5.4. One can actually show that there is an isometry between q-tame persistence modules — a purely algebraic construction —, and persistence diagrams — points above the diagonal — [CdSGO16]. Indeed, defining the *interleaving distance* as

$$\mathrm{d}_i(\mathbb{V}, \mathbb{W}) = \inf \{\delta > 0 | \mathbb{V} \text{ and } \mathbb{W} \text{ are } \delta\text{-interleaved}\},$$

we have, for all q-tame persistence modules $\mathbb{V}$ and $\mathbb{W}$,

$$\mathrm{d_b}(\mathrm{dgm}(\mathbb{V}), \mathrm{dgm}(\mathbb{W})) = \mathrm{d}_i(\mathbb{V}, \mathbb{W}).$$

5.2. **Stability for Functions.**

Although purely algebraic and rather abstract, this result is an efficient tool to easily establish concrete stability results such as the following.

Definition 5.5 (q-Tame Function)**.** Let $f : M \to \mathbb{R}$ be a real-valued functions defined on a topological space M. We say that f is *q-tame* if the sublevel sets filtrations of f induces a q-tame module at the homology level.

Proposition 5.6. *If $f : M \to \mathbb{R}$ is continuous and M is finitely triangulable (i.e. homeomorphic to a finite simplicial complex), then f is q-tame.*

Proof. Fpr simplicity, let us write $M_r = f^{-1}((-\infty, r])$, for $r \in \mathbb{R}$. For all $b < c$, we must show that $H(M_b) \to H(M_c)$ has finite rank. Begin with any finite triangulation of M, and subdivide it repeatedly until no simplex meets both $f^{-1}(b)$ and $f^{-1}(c)$. If we define $\mathcal{K}$ to be the union of the closed simplices which meet M_b, then we have

$$M_b \subset \mathcal{K} \subset M_c,$$

and hence the factorization

$$H(M_b) \to H(\mathcal{K}) \to H(M_c).$$

Since $\mathcal{K}$ is (a geometric realization of) a finite simplicial complex, $H(\mathcal{K})$ is finite dimensional and so $H(M_b) \to H(M_c)$ has finite rank. □

Theorem 5.7. *Let $f, g : M \to \mathbb{R}$ be q-tame. Then for any integer k,*

$$\mathrm{d_b}(\mathrm{dgm}_k(f), \mathrm{dgm}_k(g)) \leqslant \|f - g\|_\infty = \sup_{x \in M} |f(x) - g(x)|$$

where $\mathrm{dgm}_k(f)$ *(resp.* $\mathrm{dgm}_k(g)$*) is the persistence diagram of the persistence module $(H_k(f^{-1}(-\infty, r]))|r \in \mathbb{R})$ (resp. $(H_k(g^{-1}(-\infty, r]))|r \in \mathbb{R})$) where the linear maps are the one induced by the canonical inclusion maps between sublevel sets.*

Proof. Denoting $\delta = \|f - g\|_\infty$ we have that for any $r \in \mathbb{R}$, $f^{-1}(-\infty, r]) \subset g^{-1}(-\infty, r + \delta])$ and $g^{-1}(-\infty, r]) \subset f^{-1}(-\infty, r + \delta])$. This interleaving between the sublevel sets of f induces a δ-interleaving between the persistence modules at the homology level and the result follows from the direct application of Theorem 5.3. □

5.3. **Stability for Spaces.**

It sometimes occurs in that one has to compare data sets that are not sampled from the same ambient space. Fortunately, the notion of Hausdorff distance can be generalized to the comparison of any pair of compact metric spaces, giving rise to the notion of *Gromov-Hausdorff distance.*

Two compact metric spaces (M_1, ρ_1) and (M_2, ρ_2) are *isometric* if there exists a bijection $\phi : M_1 \to M_2$ that preserves distances, i.e. $\rho_2(\phi(x), \phi(y)) = \rho_1(x, y)$ for any $x, y \in M_1$. The Gromov-Hausdorff distance measures how far two metric space are from being isometric.

Definition 5.8. The Gromov-Haudorff distance $\mathrm{d_{GH}}(M_1, M_2)$ between two compact metric spaces is the infimum of the real numbers $r \geqslant 0$ such that there exists a metric space (M, ρ) and two compact subspaces $C_1, C_2 \subset M$ that are isometric to M_1 and M_2 and such that $\mathrm{d_H}(C_1, C_2) \leqslant r$.

Theorem 5.3 also implies a stability result for the persistence diagrams of filtrations built on top of data.

Theorem 5.9. *Let $\mathbb{X}$ and $\mathbb{Y}$ be two compact metric spaces and let* $\mathrm{Filt}(\mathbb{X})$ *and* $\mathrm{Filt}(\mathbb{Y})$ *be the Vietoris-Rips of Čech filtrations built on top $\mathbb{X}$ and $\mathbb{Y}$. Then*

$$\mathrm{d_b}\left(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X})), \mathrm{dgm}(\mathrm{Filt}(\mathbb{Y}))\right) \leqslant 2\, \mathrm{d_{GH}}(\mathbb{X}, \mathbb{Y}),$$

where $\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}))$ *and* $\mathrm{dgm}(\mathrm{Filt}(\mathbb{Y}))$ *denote the persistence diagram of the filtrations* $\mathrm{Filt}(\mathbb{X})$ *and* $\mathrm{Filt}(\mathbb{X})$.

Proof. See [CdSO14, Theorem 5.2]. □

Remark 5.10. (i) This bound is worst-case tight. Indeed, take $\mathbb{X} = \{0, 1\} \subset \mathbb{R}$ and $\mathbb{Y} = \{0, 1 + 2\varepsilon\}$, for $\varepsilon > 0$ (see Figure 6a). Then $\mathrm{d_{GH}}(\mathbb{X}, \mathbb{Y}) = \varepsilon$, $\mathrm{dgm}_0(\mathrm{Filt}(\mathbb{X})) = \{(0, \infty), (0, 1)\}$ and $\mathrm{dgm}_0(\mathrm{Filt}(\mathbb{Y})) = \{(0, \infty), (0, 1 + 2\varepsilon)\}$, so that

$$\mathrm{d_b}\left(\mathrm{dgm}_0(\mathrm{Filt}(\mathbb{X})), \mathrm{dgm}_0(\mathrm{Filt}(\mathbb{Y}))\right) = \varepsilon = 2\, \mathrm{d_{GH}}(\mathbb{X}, \mathbb{Y}).$$

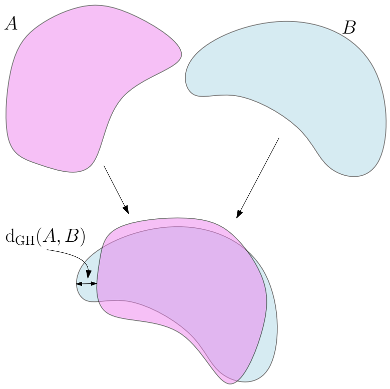


FIGURE 5. The Gromov-Hausdorff distance between $A, B \subset \mathbb{R}^2$. A can been rotated — this is an isometric embedding of A in the plane — to reduce its Hausdorff distance to B. As a consequence, $\mathrm{d}_{\mathrm{GH}}(A,B) < \mathrm{d}_{\mathrm{H}}(A,B)$ in this case.

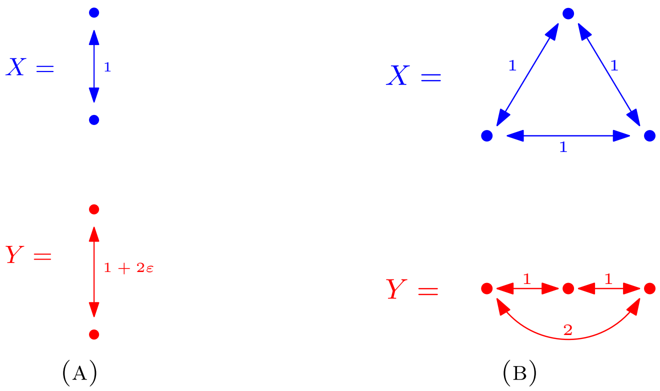


FIGURE 6. Discussion on the tightness of Theorem 5.9.

(ii) In general, this is only an upper bound. Indeed, write

$$\mathbb{X} = \{(0,0), (1,0), (1/2, \sqrt{3}/2)\} \subset \mathbb{R}^2$$

and $\mathbb{Y} = \{-1, 0, 1\}$ (see Figure 6b). Then $\mathrm{d}_{\mathrm{GH}}(\mathbb{X}, \mathbb{Y}) = 1/2$, while

$$\mathrm{dgm}_0(\mathrm{Filt}(\mathbb{X})) = \{(0,\infty), (0,1), (0,1)\} = \mathrm{dgm}_0(\mathrm{Filt}(\mathbb{Y})),$$

so that

$$\mathrm{d}_{\mathrm{b}}\left(\mathrm{dgm}_0(\mathrm{Filt}(\mathbb{X})), \mathrm{dgm}_0(\mathrm{Filt}(\mathbb{Y}))\right) = 0.$$

(iii) The proofs never use the triangle inequality! The previous approach and results easily extend to other settings like, e.g. spaces endowed with a similarity measure.

(iv) As we already noticed in Example 2.3, the persistence diagrams can be interpreted as multiscale topological features of $\mathbb{X}$ and $\mathbb{Y}$. In addition, Theorem 5.9 tells us that these features are robust with respect to perturbations of the data in the Gromov-Hausdorff metric.

6. Rates of Convergence for Random Point Clouds

Persistence homology by itself does not take into account the random nature of data and the intrinsic variability of the topological quantity they infer. We now present a statistical approach to persistent homology, which means that we consider data as generated from an unknown distribution.

6.1. Minimax Upper Bound.

Assume that we observe an i.i.d. n-sample $\mathbb{X}_n = \{X_1, \dots, X_n\}$ in a metric space (M, ρ) drawn from an unknown probability measure μ, whose support is a compact set denoted by $\mathbb{X}_\mu$.

Let $\mathrm{Filt}(\mathbb{X}_\mu)$ and $\mathrm{Filt}(\widehat{\mathbb{X}})$ be two filtrations defined on $\mathbb{X}_\mu$ and $\widehat{\mathbb{X}}$. Starting from Theorem 5.9, a natural strategy for estimating the persistent homology of $\mathrm{Filt}(\mathbb{X}_\mu)$ is to consider that of $\mathrm{Filt}(\widehat{\mathbb{X}})$, where $\widehat{\mathbb{X}}$ is an estimator of $\mathbb{X}_\mu$, meaning that $\mathrm{d}_{\mathrm{GH}}(\mathbb{X}_\mu, \widehat{\mathbb{X}})$ is small.

Remark 6.1. Note that in some cases the space M can be unknown and the observations $X_1 \dots, X_n$ are then only known through their pairwise distances $(\rho(X_i, X_j))_{1 \leqslant i,j \leqslant n}$. The use of the Gromov-Hausdorff distance allows us to consider this set of observations as an abstract metric space of cardinality n, independently of the way it is embedded in M.

Definition 6.2 ((a,b)-Standard Measure)**.** The distribution μ is said to be (a,b)*-standard* if for all $x \in \mathrm{supp}(\mu)$ and all $r \geqslant 0$,

$$\mu\left(\mathrm{B}(x,r)\right) \geqslant \min(ar^b, 1).$$

The finite set $\mathbb{X}_n := \{X_1, \dots, X_n\}$ is a natural estimator of the support $\mathbb{X}_\mu$. In several contexts discussed in the following, $\mathbb{X}_n$ shows optimal rates of convergence to $\mathbb{X}_\mu$ with respect to the Hausdorff distance. A slight variant of this assumption has already been used in the previous lessons.

Definition 6.3 (Statistical Model)**.** We let $\mathcal{P}_{M,a,b}$ denote the set of Borel probability distributions μ over (M, ρ) such that

– $\mathbb{X}_\mu = \mathrm{supp}\,\mu$ is compact;
– μ is (a,b)-standard.

The following result gives an upper bound for the rate of convergence of persistence diagrams for (a,b)-standard measures.

Theorem 6.4. *If μ is (a,b)-standard on (M,ρ), then :*

(i) For all $\varepsilon > 0$,

$$\mathbb{P}\left(\mathrm{d_b}\left(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n))\right) > \varepsilon\right) \leqslant \min\left(\frac{2^b}{a\varepsilon^b}\exp(-na\varepsilon^b), 1\right).$$

(ii) For n large enough,

$$\sup_{\mu \in \mathcal{P}_{M,a,b}} \mathbb{E}_{\mu^n}\left[\mathrm{d_b}(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n)))\right] \leqslant C_{a,b}\left(\frac{\log n}{n}\right)^{1/b}.$$

Proof. To get (i), we apply the stability of persistence for spaces Theorem 5.9 to get

$$\begin{aligned}\mathbb{P}\left(\mathrm{d_b}\left(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n))\right) > \varepsilon\right) &\leqslant \mathbb{P}\left(\mathrm{d_{GH}}\left(\mathbb{X}_\mu, \mathbb{X}_n\right) > \varepsilon/2\right) \\ &\leqslant \mathbb{P}\left(\mathrm{d_H}\left(\mathbb{X}_\mu, \mathbb{X}_n\right) > \varepsilon/2\right) \\ &\leqslant \min\left(\frac{2^b}{a\varepsilon^b}\exp(-na\varepsilon^b), 1\right),\end{aligned}$$

where the last inequality follows from a packing argument, as already detailed in the previous lessons. Moving to the proof of (ii), we use Fubini's theorem to write

$$\begin{aligned}&\mathbb{E}_{\mu^n}\left[\mathrm{d_b}(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n)))\right] \\ &\quad = \int_0^\infty \mathbb{P}\left(\mathrm{d_b}(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n))) > \varepsilon\right) \mathrm{d}\varepsilon.\end{aligned}$$

Let $\varepsilon_n = 4\left(\frac{\log n}{an}\right)^{1/b}$. By bounding the probability inside this integral by one on $[0, \varepsilon_n]$, we get

$$\begin{aligned}&\mathbb{E}_{\mu^n}\left[\mathrm{d_b}(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathrm{dgm}(\mathrm{Filt}(\widehat{\mathbb{X}}_n)))\right] \\ &\quad \leqslant \varepsilon_n + \int_{\varepsilon_n}^\infty \frac{8^b}{a}\varepsilon^{-b}\exp(-na\varepsilon^b/4^b)\mathrm{d}\varepsilon \\ &\quad \leqslant \varepsilon_n + \frac{4n2^b}{b}(na)^{-1/b}\int_{\log n}^\infty u^{1/b-2}\exp(-u)\mathrm{d}u.\end{aligned}$$

We now distinguish two cases.

If $b \geqslant \frac{1}{2}$: then $u^{1/b-2} \leqslant (\log n)^{1/b-2}$ for all $u \geqslant \log n$ and then

$$\begin{aligned}\mathbb{E}\left[\mathrm{d_b}(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathrm{dgm}(\mathrm{Filt}(\widehat{\mathbb{X}}_n)))\right] &\leqslant \varepsilon_n + 4\frac{2^b}{b}\left(\frac{\log n}{n}\right)^{1/b}(\log n)^{-2} \\ &\leqslant C_{a,b}\left(\frac{\log n}{n}\right)^{1/b},\end{aligned}$$

where $C_{a,b}$ only depends only a and b.

If $0 < b < \frac{1}{2}$: we let $p := \lfloor 1/b \rfloor$ and $u_n := \log n$. Using iterated integrations by parts yields

$$\begin{aligned}&\int_{u_n}^\infty u^{1/b-2}\exp(-u)\mathrm{d}u \\ &\quad = u_n^{1/b-2}\exp(u_n) + (\frac{1}{b} - 2)u_n^{1/b-3}\exp(u_n) + \cdots + \\ &\quad + \prod_{i=2}^{p}\left(\frac{1}{b} - i\right)u_n^{1/b-p}\exp(u_n) + \int_{\log n}^\infty u^{1/b-p-1}\exp(-u)\mathrm{d}u \\ &\quad \leqslant C'_{a,b}\frac{(\log n)^{1/b-2}}{n},\end{aligned}$$

where $C'_{a,b}$ only depends only a and b.

Thus, the expected loss bound holds for all $b > 0$, yielding the result. □

6.2. **Minimax Lower Bound.** Let us recall Le Cam's Lemma.

LEMMA 6.5 (Le Cam). *Let $\mathcal{Q}$ be a set of probability distributions, and $\theta : \mathcal{Q} \to \Theta$ be a parameter of interest, where (Θ, ℓ) is a metric space.*
Then for all $Q, Q' \in \mathcal{Q}$,

$$\inf_{\hat\theta} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n} \ell\big(\theta(Q), \hat\theta_n\big) \geqslant \frac{1}{2} \ell\left(\theta(Q), \theta(Q')\right) \left(1 - \mathrm{TV}(Q, Q')\right)^n,$$

where $\hat\theta_n = \hat\theta_n(X_1, \ldots, X_n)$ ranges among all the measurable maps $\hat\theta_n : \mathcal{X}^n \to \Theta$ based on an i.i.d. n-sample.

THEOREM 6.6. *Assume that there exists a non isolated point x in M and consider any sequence $(x_n)_n \in (M \setminus \{x\})^{\mathbb{N}}$ such that $\rho(x, x_n) \leqslant (an)^{-1/b}$. Then for all estimator $\widehat{\mathrm{dgm}}_n = \widehat{\mathrm{dgm}}_n(X_1, \ldots, X_n)$,*

$$\liminf_{n \to \infty} \rho(x, x_n)^{-1} \sup_{\mu \in \mathcal{P}_{M,a,b}} \mathbb{E}_{\mu^n} \left[\mathrm{d_b}(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \widehat{\mathrm{dgm}}_n)\right] \geqslant e^{-1}/4.$$

Remark 6.7. Consequently, the estimator $\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n))$ is minimax optimal on the space $\mathcal{P}_{M,a,b}$ up to a logarithmic term as soon as we can find a non-isolated point in M and a sequence (x_n) in M such that $\rho(x_n, x) \sim (an)^{-1/b}$. This is obviously the case for the Euclidean space $\mathbb{R}^d$.

Proof. We will apply Le Cam's lemma with model $\mathcal{Q} = \mathcal{P}_{M,a,b}$, parameter of interest $\theta(\mu) = \mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_\mu))$ in the space Θ of persistence diagrams of q-tame modules endowed with distance $\ell = \mathrm{d_b}$.

To prove the lower bound, it will be sufficient to consider two Dirac distributions. We let $\mu_0 = \delta_x$ be the Dirac distribution on $\mathbb{X}_0 := \{x\}$. It is clear that $\mu_0 \in \mathcal{P}_{M,a,b}$. Let $\mu_{1,n}$ be the distribution $\frac{1}{n}\delta_{x_n} + (1 - \frac{1}{n})\mu_0$. The support of $\mu_{1,n}$ is denoted $\mathbb{X}_{1,n} := \{x, x_n\}$. Note that for all $n \geqslant 2$ and $r \leqslant \rho(x, x_n)$,

$$\mu_{1,n}\left(\mathrm{B}(x, r)\right) = 1 - \frac{1}{n} \geqslant \frac{1}{2} \geqslant \frac{1}{2\rho(x, x_n)^b} r^b \geqslant ar^b$$

and

$$\mu_{1,n}\left(\mathrm{B}(x_n, r)\right) = \frac{1}{n} = \frac{1}{n\rho(x, x_n)^b} r^b \geqslant ar^b.$$

Moreover, for $r > \rho(x, x_n)$, $\mu_{1,n}\left(\mathrm{B}(x, r)\right) = \mu_{1,n}\left(\mathrm{B}(x_n, r)\right) = 1$. Thus, for all $r > 0$ and $x \in \mathbb{X}_{1,n}$,

$$\mu_{1,n}\left(\mathrm{B}(x, r)\right) \geqslant \min\{ar^b, 1\},$$

meaning that $\mu_{1,n}$ belongs to $\mathcal{P}_{M,a,b}$.

The probability measure μ_0 is absolutely continuous with respect to $\mu_{1,n}$ and the density of μ_0 with respect to $\mu_{1,n}$ is $p_{0,n} := \frac{n}{n-1}\mathbb{1}_{\{x\}}$. Then

$$\mathrm{TV}(\mu_0, \mu_{1,n}) = \frac{1}{2}\int_M \left|1 - \frac{n}{n-1}\mathbb{1}_{\{x\}}\right| \mathrm{d}\mu_{1,n} = \frac{1}{n},$$

so that $(1 - \mathrm{TV}(\mu_0, \mu_{1,n}))^n = \left(1 - \frac{1}{n}\right)^n \to e^{-1}$ as n goes to infinity.

It remains to compute $\mathrm{d_b}(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_0)), \mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_{1,n})))$. For both $\mathbb{X}_0$ and $\mathbb{X}_{1,n}$, notice that the diagrams induced by the Rips and Čech filtrations are equal and that these diagrams are non-trivial only for the 0-dimensional homology. Furthermore, $\mathrm{dgm}_0\left(\mathrm{Filt}(\mathbb{X}_0)\right)$ is the singleton $\{(0, +\infty)\}$. On the other hand, $\mathrm{dgm}_0\left(\mathrm{Filt}(\mathbb{X}_{1,n})\right) = \{(0, \infty), (0, \rho(x, x_n))\}$. Thus,

$$\begin{aligned}\mathrm{d_b}(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_0)), \mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_{1,n}))) &= \min_{p\in\Delta} \|p - (0, \rho(x, x_n))\|_\infty \\ &= \frac{\rho(x, x_n)}{2}.\end{aligned}$$

The proof is then complete using Le Cam's lemma (Lemma 6.5). □

7. Persistence Landscapes

Persistence landscapes have been introduced in [Bub15] as an alternative representation of persistence diagrams. This approach aims at representing the topological information encoded in persistence diagrams as elements of an Hilbert space, for which statistical learning methods can be directly applied.

7.1. **Construction.**

The persistence landscape is a collection of continuous, piecewise linear functions $\lambda\colon \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ that summarizes a persistence diagram dgm (see Figure 7). The landscape is defined by considering the set of tent functions at each point $p = (x, y) = \left(\frac{\alpha_{\text{birth}}+\alpha_{\text{death}}}{2}, \frac{\alpha_{\text{death}}-\alpha_{\text{birth}}}{2}\right)$ representing a birth-death pair $(\alpha_{\text{birth}}, \alpha_{\text{death}}) \in \mathrm{dgm}$ as follows:

$$\begin{aligned}\Lambda_p(t) &= \begin{cases} t - x + y & t \in [x - y, x] \\ x + y - t & t \in (x, x + y] \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} t - \alpha_{\text{birth}} & t \in \left[\alpha_{\text{birth}}, \frac{\alpha_{\text{birth}}+\alpha_{\text{death}}}{2}\right] \\ \alpha_{\text{death}} - t & t \in \left(\frac{\alpha_{\text{birth}}+\alpha_{\text{death}}}{2}, \alpha_{\text{death}}\right] \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

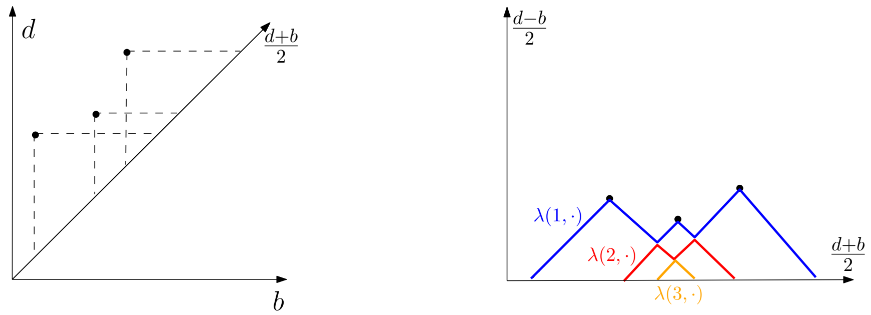


Figure 7. An example of persistence landscape (right) associated to a persistence diagram (left). The first landscape is in blue, the second one in red and the last one in orange. All the other landscapes are zero.

The persistence landscape of dgm is a summary of the arrangement of the tents display obtained by overlaying the graphs of the functions $\{\Lambda_p\}_{p\in\mathrm{dgm}}$.

Definition 7.1 (Landscape of a Diagram)**.** The persistence landscape of a diagram dgm is the collection of functions, indexed by $k \in \mathbb{N}$, and defined by

$$\lambda_{\mathrm{dgm}}(k,t) = \operatorname*{kmax}_{p \in \mathrm{dgm}} \Lambda_p(t), \quad t \in [0,T],$$

where kmax is the kth largest value in the set; in particular, 1max is the usual maximum function.

Given $k \in \mathbb{N}$, the function $\lambda_{\mathrm{dgm}}(k,.) : \mathbb{R} \to \mathbb{R}$ is called the *k-th landscape* of dgm. It is not difficult to see that the map that associate to each persistence diagram its corresponding landscape is injective. In other words, formally no information is lost when a persistence diagram is represented through its persistence landscape.

The advantage of the persistence landscape representation is two-fold:

(i) Persistence diagrams are represented as elements of a function space, opening the door to the use of a broad variety of statistical and data analysis tools for further processing of topological features.

(ii) Second, and fundamental from a theoretical perspective, the persistence landscapes share the same stability properties as persistence diagrams (see Section 5).

Proposition 7.2 (Basic Properties of Landscapes). *For all $k \geqslant 0$,*

(i) $\lambda_{\mathrm{dgm}}(k,\cdot) \geqslant \lambda_{\mathrm{dgm}}(k+1,\cdot) \geqslant 0$,

(ii) $\lambda_{\mathrm{dgm}}(k,\cdot)$ *is* 1*-Lipschitz.*

Proof. See [Bub15, Lemma 4]. □

7.2. **Stability.** From the definition of persistence landscape, we immediately observe that $\lambda(k,\cdot)$ is one-Lipschitz and thus similar stability properties are satisfied for the landscapes as for persistence diagrams.

Theorem 7.3 (Stability of Landscapes). *Let* dgm_1 *and* dgm_2 *be two q-tame diagrams. Then for all $k \geqslant 0$,*

$$\left\| \lambda_{\mathrm{dgm}_1}(k,\cdot) - \lambda_{\mathrm{dgm}_2}(k,\cdot) \right\|_\infty \leqslant \mathrm{d_b}(\mathrm{dgm}_1, \mathrm{dgm}_2).$$

Proof. See [Bub15, Theorem 17]. □

Remark 7.4. In particular, Theorem 7.3 allows to derive a stability result for landscapes associated to:

(i) filtrations of functions, from Theorem 5.7;

(ii) Rips and Čech filtrations of a metric space, from Theorem 5.9.

7.3. **Central Tendency for Persistent Homology.** The space of persistence diagrams being not an Hilbert space, the definition of a *mean persistence diagram* is not obvious and unique. One first approach to define a central tendency in this context is to define a Fréchet mean in this context. Indeed it has been proved in [TMMH14] that the space of persistence diagrams is a Polish space. However they are may not be unique and there are very difficult to compute in practice. To overcome the problem of computational costs, sampling strategies can be proposed to compute topological signatures based on persistence landscapes. Given a large point cloud, the idea is to extract many subsamples, to compute the landscape for each subsample and then to combine the information.

We assume that the diameter of M is finite and upper bounded by $\frac{T}{2}$, where T is the same constant as in the definition of persistence landscapes in Section 7.1. For ease of exposition, we focus on the case $k = 1$, and set $\lambda(t) = \lambda(1, t)$. However, the results we present in this section hold for $k > 1$.

For any positive integer m, let $X = \{x_1, \cdots, x_m\} \subset \mathbb{X}_\mu$ be a sample of m points from μ. The corresponding persistence landscape is λ_X and we denote by Ψ_μ^m the measure induced by $\mu^{\otimes m}$ on the space of persistence landscapes. Note that the persistence landscape λ_X can be seen as a single draw from the measure Ψ_μ^m. The point-wise expectations of the (random) persistence landscape under this measure is defined by $\mathbb{E}_{\Psi_\mu^m}[\lambda_X(t)], t \in [0, T]$. The average landscape $\mathbb{E}_{\Psi_\mu^m}[\lambda_X]$ has a natural empirical counterpart, which can be used as its unbiased estimator. Let $S_1^m, \ldots, S_\ell^m$ be ℓ independent samples of size m from $\mu^{\otimes m}$. We define the empirical average landscape as

$$\overline{\lambda_\ell^m}(t) = \frac{1}{b}\sum_{i=1}^{b} \lambda_{S_i^m}(t), \qquad \text{for all } t \in [0, T],$$

and propose to use $\overline{\lambda_\ell^m}$ to estimate $\lambda_{\mathbb{X}_\mu}$.

Remark 7.5. (i) Note that computing the persistent homology of $\mathbb{X}_n$ is $O(\exp(n))$, whereas computing the average landscape is $O(b\exp(m))$.

(ii) Another motivation for this subsampling approach is that it can be also applied when μ is a discrete measure with support $\mathbb{X}_N = \{x_1, \ldots, x_N\} \subset M$. This framework can be very common in practice, when a continuous (but unknown measure) is approximated by a discrete uniform measure μ_N on $\mathbb{X}_N$.

The average landscape $\mathbb{E}_{\Psi_\mu^m}[\lambda_X]$ is an interesting quantity on its own, since it carries some stable topological information about the underlying measure μ, from which the data are generated. In particular, we can compare the average landscapes corresponding to two measures that are close to each other in the Wasserstein metric. The average behavior of the landscapes of sets of m points sampled according to any measure μ is stable with respect to the Wasserstein distance:

Theorem 7.6. *Let $X \sim \mu^{\otimes m}$ and $Y \sim \nu^{\otimes m}$, where μ and ν are two probability measures on M. For any $p \geqslant 1$ we have*

$$\left\|\mathbb{E}_{\Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y]\right\|_\infty \leqslant 2m^{\frac{1}{p}}\, \mathrm{W}_p(\mu, \nu),$$

where W_p *stands for the pth* Wasserstein distance *on M, defined by*

$$\mathrm{W}_p(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left(\int_{M\times M} \rho(x, y)^p \pi(\mathrm{d}x, \mathrm{d}y)\right)^{\frac{1}{p}},$$

where $\Pi(\mu, \nu)$ is the set of probability measures on $M \times M$ with marginal distributions μ and ν,

Proof. See [CFL+15, Theorem 5] □

Remark 7.7. (i) For measures that are not defined on the same metric space, the inequality of Theorem 7.6 can be extended, to the condition of using the so-called *Gromov-Wasserstein metric*, and writes as

$$\left\|\mathbb{E}_{\Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y]\right\|_\infty \leqslant 2m^{\frac{1}{p}} GW_{\rho,p}(\mu,\nu).$$

(ii) The result of Theorem 7.6 is useful for two reasons. First, it tells us that for a fixed m, the expected "topological behavior" of a set of m points carries some stable information about the underlying measure from which the data are generated. Second, it provides a lower bound for the Wasserstein distance between two measures, based on the topological signature of samples of m points.

8. Further Sources

These notes mainly follow [CM17], [BCY18] and [CGLM15].

References

[BCY18] Jean-Daniel Boissonnat, Frédéric Chazal, and Mariette Yvinec. *Geometric and Topological Inference.* Cambridge University Press, 2018. Cambridge Texts in Applied Mathematics.

[Bub15] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16:77–102, 2015.

[CdSGO16] Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. *The structure and stability of persistence modules.* SpringerBriefs in Mathematics. Springer, [Cham], 2016.

[CdSO14] Frédéric Chazal, Vin de Silva, and Steve Oudot. Persistence stability for geometric complexes. *Geom. Dedicata*, 173:193–214, 2014.

[CFL$^+$15] Frederic Chazal, Brittany Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Subsampling methods for persistent homology. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2143–2151, Lille, France, 07–09 Jul 2015. PMLR.

[CGLM15] Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Convergence rates for persistence diagram estimation in topological data analysis. *J. Mach. Learn. Res.*, 16:3603–3635, 2015.

[CM17] Frédéric Chazal and Bertrand Michel. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. *arXiv e-prints*, page arXiv:1710.04019, Oct 2017.

[TMMH14] Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet means for distributions of persistence diagrams. *Discrete Comput. Geom.*, 52(1):44–70, 2014.