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

Lecture 7: Topological descriptors for geometric data

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.

7.1 A distance between metric spaces: the Gromov-Hausdorff distance

We focus now on applying persistent homology to geometric data: basically the input data can be a 3D-shape, a point cloud, etc. The general framework to encode such items is to consider our data to be *metric spaces*. We want a notion of distance between our input data, that is a distance between metric spaces. This will be the Gromov-Hausdorff distance.

We first recall the definition of the Hausdorff distance, introduced in previous lecture, which is a distance between sets included in a same metric space.

Definition 1. *Let (Z, d_Z) be a metric space, and let $X, Y \subset Z$.*

$$d_H^Z(X,Y) := \max \left\{ \sup_{x \in X} \inf_{y \in Y} d_Z(x,y), \sup_{y \in Y} \inf_{x \in X} d_Z(x,y) \right\} \tag{7.1}$$

We can now define the Gromov-Hausdorff distance:

Definition 2 (Gromov-Hausdorff distance)**.** *Let (X, d_X) and (Y, d_Y) be two metric spaces.*

$$d_{\mathrm{GH}}(X,Y) := \inf_{(Z,d_Z),\gamma_X,\gamma_Y} d_H^Z(\gamma_X(X), \gamma_Y(Y)) \tag{7.2}$$

where the infimum is taken under the constraints:

- (Z, d_Z) *is a metric space*
- γ_X *(resp γ_Y) is an isometry between X and Z (resp Y and Z).*

Interpretation: Given two metric spaces, we are looking for all common (isometric) embedding of these spaces, and among all possible such embeddings, we are interested in the one that minimizes the Hausdorff distance between the two spaces.

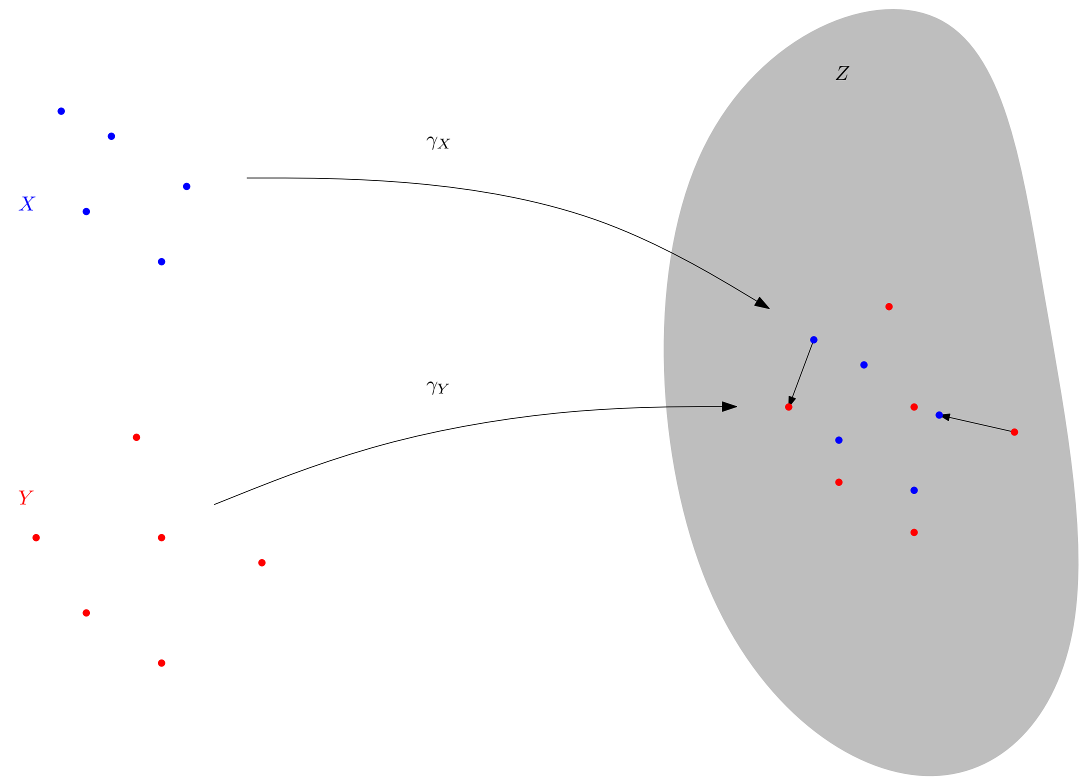

Figure 7.1: Schematic illustration of the Gromov-Hausdorff distance.

7.2 Topological descriptors

Consider (X, d_X) a compact metric space. Informally, a *descriptor* (or a *signature*) on (X, d_X) is a way to summarize X. The idea is then to use descriptors instead of initial objects to perform statistical and learning tasks. Basically, a descriptor should be (in an ideal world, not exhaustive list):

- easy to compute,
- easy to compare,
- stable (if two initial objects are similar, corresponding descriptors should be similar),
- inverse-stable (if two descriptors are similar, input objects should be similar),
- interpretable,
- useful in applications.

For example, given a set of real numbers, their arithmetic mean is a descriptor of this set. It is easy to compute, easy to compare (we even get an order so we can rank things!), it's somewhat stable (details are skipped), but not inverse-stable (two very different sets could have the same arithmetic mean). In the context of geometric data, there are many descriptors (cf slides page 3). In this lecture, we see how persistence diagrams can be used as (topological) descriptors for geometric data.

7.2.1 Global signatures

Let (X, d_X) be a compact metric space. We recall the definition of the Čech filtration (adapted to a metric space):

Definition 3 (Čech (or nerve) filtration).

$$\mathcal{C}(X, d_X) = (C_t(X, d_X))_{t \geq 0},$$
$$\sigma = \{x_0 \dots x_r\} \subset X \in C_t(X, d_X) \Leftrightarrow \bigcap_{i=0}^{r} B(x_i, t) \neq \emptyset.$$

We have the nerve theorem:

Theorem 1 (Nerve). *Let $X \subset (Z, d_Z)$ (equipped with the endowed metric). Suppose that $\forall \sigma \subset X$ finite, $\bigcap_{x \in \sigma} B_Z(x, t)$ is either empty or contractible. Then, $C_t(X, d_Z)$ is homotopy equivalent to $\bigcup_{x \in X} B_Z(x, t)$.*

Interpretation: The Čech filtration is a combinatorial proxy for the union of ball with radius t centered in X.

Definition 4 (Vietoris-Rips filtration).

$$\mathcal{R}(X, d_X) := (R_t(X, d_X))_{t \geq 0}$$
$$\sigma = \{x_0 \dots x_r\} \in R_t(X, d_X) \Leftrightarrow \underbrace{\max_{x_i, x_j \in \sigma} \|x_i - x_j\|}_{\mathrm{diam}(\sigma)} \leq t$$

Remark: The Rips filtration is easier to compute than the Čech one.

Proposition 1. *For any (X, d_X) compact,*

$$\forall t \in \mathbb{R}, R_t(X, d_X) \subseteq C_t(X, d_X) \subseteq R_{2t}(X, d_X)$$

Proof. Exercise. □

Proposition 2. *If $(X, d_X) = (\mathbb{R}^d, \ell^\infty)$, then $\mathcal{C}(X, d_X) = \mathcal{R}(X, d_X)$.*

Theorem 2 (Stability theorem). *For (X, d_X) and (Y, d_Y) two compact metric spaces,*

$$d_B^\infty(\mathrm{Dgm}(\mathcal{R}(X, d_X)), \mathrm{Dgm}(\mathcal{R}(Y, d_Y))) \leq 2 d_{\mathrm{GH}}(X, Y) \tag{7.3}$$

Remark: This bound is tight. However, there is no converse inequality: you can find different (in Gromov-Hausdorff sense) metric spaces having same diagrams.

7.2.2 Local signatures

Definition 5. *A metric space (X, d_X) is said to be* intrinsic *if*

$$\forall x, y \in X, d_X(x, y) = \inf_{\gamma \in \Pi(x, y)} \mathcal{L}(\gamma)$$

where $\Pi(x, y) := \{\gamma : [0, 1] \to X$ continuous $, \gamma(0) = x, \gamma(1) = y\}$ is the set of path *between x and y, and*

$$\mathcal{L}(\gamma) := \inf_{n \in \mathbb{N}, 0 = t_0 \leq \dots t_n = 1} \sum_{i=0}^{n-1} d_X(\gamma(t_i, t_{i+1})$$

is the length *of γ.*

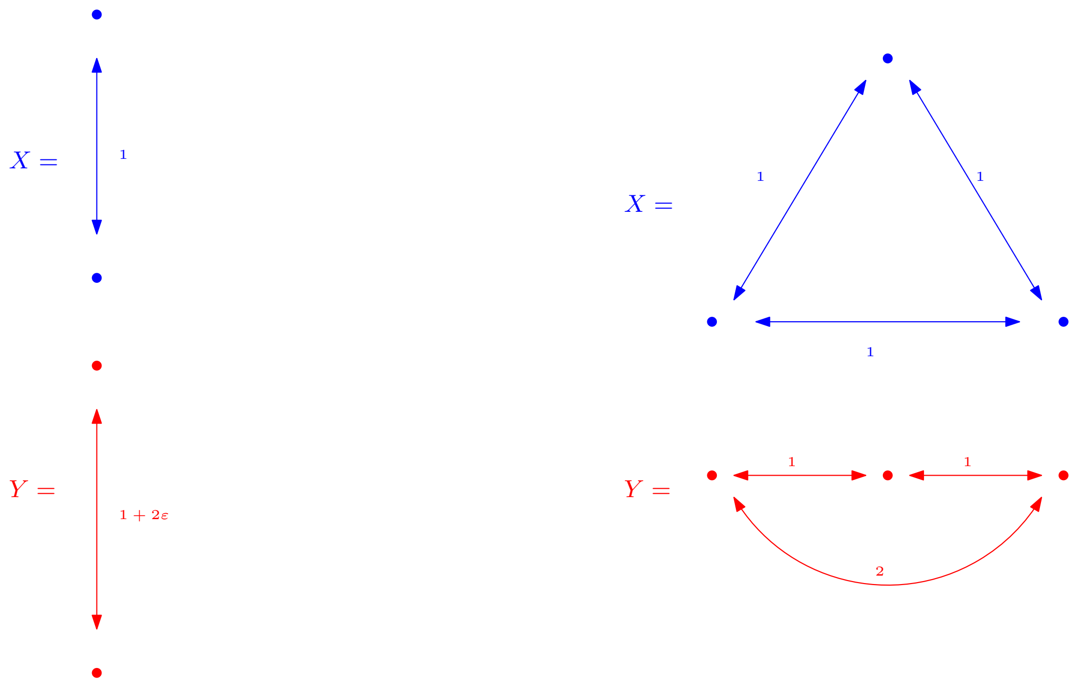

Figure 7.2: *(left)* Case of two metric spaces (point cloud) such that $d_{\mathrm{GH}} = 2\varepsilon$ and $d_B^{\infty} = \varepsilon$. *(right)* Example where $d_{\mathrm{GH}} = 1/2$ while $d_B^{\infty} = 0$.

Definition 6. *Given (X, d_X) intrinsic metric space, $x \in X$,*

$$\rho(X, x) := \sup\{r > 0 | \forall r' < r, B_X(x, r') \textit{ is convex }\}.$$

Being convex means $\forall y, z \in B_X(x, r'), \exists!$ shortest path $\in \Pi(y, z)$ and that path lies in $B_X(x, r')$.

$$\rho(X) := \inf_{x \in X} \rho(X, x)$$

Definition 7 (Gromov-Hausdorff distance between pointed spaces)**.** *Let $X, Y \subset (Z, d_Z)$, and $x_0 \in x, y_0 \in Y$. Consider the* pointed space *(X, x_0) and (Y, y_0). Their Hausdorff distance is defined as*

$$d_H^Z((X, x_0), (Y, y_0)) := \max\{d_Z(x_0, y_0), d_H^Z(X, Y)\}.$$

Consider now two metric spaces (X, d_X) and (Y, d_Y), and two points $x_0 \in x, y_0 \in Y$. The Gromov-Hausdorff distance between (X, x_0, d_X) and (Y, y_0, d_Y) is defined as

$$d_{\mathrm{GH}}((X, x_0, d_X), (Y, y_0, d_Y)) := \inf_{(Z, d_Z), \gamma_X, \gamma_Y} d_H^Z((\gamma_X(X), \gamma_X(x_0)), (\gamma_Y(Y), \gamma_Y(y_0)))$$

where the infimum is taken under the constraints:

- *(Z, d_Z) is a metric space*
- *γ_X (resp γ_Y) is an isometry between X and Z (resp Y and Z).*