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

Lecture 2: Clustering and introduction to persistence

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.

2.1 Mode-seeking

Input: A point cloud $(p_1 \dots p_n) \subset \mathbb{R}^d$.

Assumptions:

- The p_is are *sampled i.i.d* according to some unknown probability distribution μ with (unknown) density f with respect to the Lebesgue measure.
- f is *regular*, typically a Morse function: twice differentiable, finitely many critical points, non-degenerate (Hessian matrix is non-singular), all distinct criticala values.

Note: The gradient vector field $x \mapsto \nabla f(x)$ is Lipschitz continuous which implies that it can be integrated into a *gradient flow* $\Phi : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$ whose trajectories are solution of the ODE (Cauchy-Lipschitz theorem):

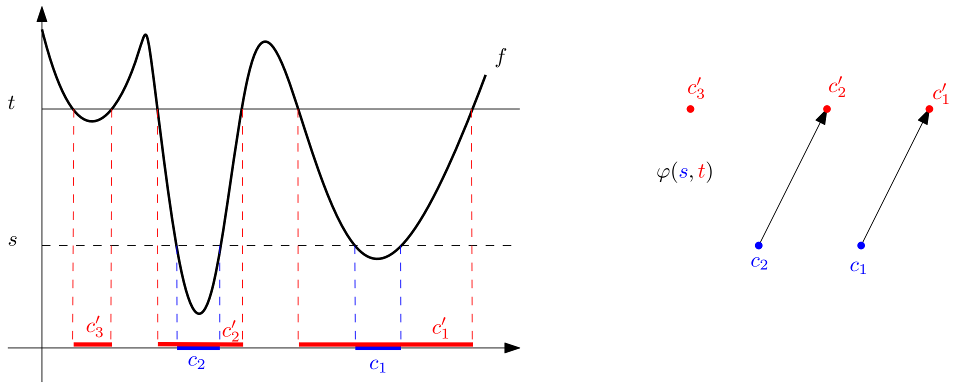
$$\begin{aligned} \gamma_x'(t) &= \nabla f \circ \gamma_x(t) \\ \gamma_x(0) &= x \end{aligned}$$

Theorem 1. *If f is Morse, then almost every point of $\mathbb{R}^d$ ends up at a maximum of f when following the gradient flow (integration of the gradient vector field) of f.*

Principle: Cluster $\mathbb{R}^d$ by the ascending regions ($\mathrm{Asc}(p) = \{x \in \mathbb{R}^d | p \in Im\gamma_x\}$) of the peaks of f. In practice, it can be simulated by *Hill-climbing algorithm*.

Limitations: In practice, we don't have access to f, and produce with our observations $p_1 \dots p_n$ an estimator of f noted $\hat{f}_n$. This estimator may be noisy, which makes appear many (irrelevant) local-maxima, all of them identified as an individual cluster.

(One) Solution: Compute the *persistence* (see below for an introduction) of local maxima and filter those with low persistence (viewed as *topological noise*).



2.2 Degree-0 persistence (Size theory)

2.2.1 Definition

Input: $f : X \to \mathbb{R}$.
Idea: In order to understand (degree-0) topology of f, we want to look at the (path-)connected components of the *excursion* sets induced by f (called a *filtration*):

- *sublevel sets*: $f^{-1}((-\infty, t])$
- *superlevel sets*: $f^{-1}([t, +\infty))$

Assumption: f is *tame*, i.e. every excursion set has finitely many (path-)connected components (cc). We define:

$$F(t) := f^{-1}((-\infty, t])$$
$$\pi_0 F(t) := \{\text{cc of } F(t)\}$$

Observation: For any two reals s, t with $s \leq t$, we have that $\forall c \in \pi_0 F(s), \exists! c' \in \pi_0 F(t)$ satisfying $c \cap c' \neq \emptyset$ (in fact, $c \subset c'$). This is because the connected components frow with t and are, by definition, pairwise disjoint. Thus, one can define an *induced map* $\varphi(s,t) : \pi_0 F(s) \to \pi_0 F(t)$ that tells where each connected component of $F(s)$ "goes" in $F(t)$.

Definition 1. *Given $t \in \mathbb{R}$, and $c \in \pi_0 F(t)$, we define*

- Birth time: $b(c) := \inf \{s \leq t | c \in \text{Im}\varphi(s,t)\}$
- Death time: $d(c) := \sup \left\{u \geq t | \forall c' \in \varphi(t,u)^{-1}(\{\varphi(t,u)(c)\}), b(c') \geq b(c)\right\}$

The interval $[b(c), d(c)]$ is the bar corresponding to c in the barcode *of f. Formally,*

$$\text{Barcode}(f) := \{[b(c), d(c)] | c \in \mathcal{C}\} \tag{2.1}$$

where

$$\mathcal{C} := \{c | c \in \pi_0 F(t) \textit{ for some } t \in \mathbb{R}\} / \sim$$

where for $t \leq u$ and $c \in \pi_0 F(t), c' \in \pi_0 F(u)$ (+ symmetry if $u \leq t$),

$$c \sim c' \Leftrightarrow \varphi(t,u)(c) = c' \textit{ and } u \leq d(c)$$

Definition 2. *The* persistence diagram *of f (associated with dimension 0) is the* multiset[1]*:*

$$\text{Dgm}(f) := \left\{(b(c), d(c)) \in \mathbb{R}^2 | c \in \mathcal{C}\right\} \tag{2.2}$$

Theorem 2 (Stability theorem)**.** *For any tame functions $f, g : X \to \mathbb{R}$,*

$$d_B^\infty(\text{Dgm}(f), \text{Dgm}(g)) \leq \|f - g\|_\infty \tag{2.3}$$

[1] It means that some points can be repeted.

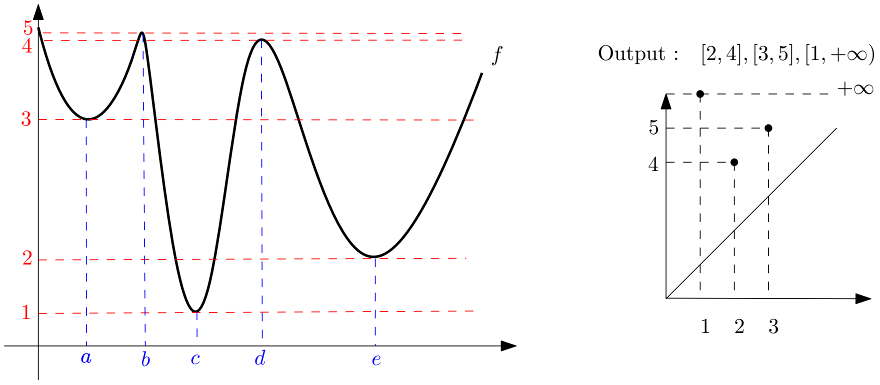


Figure 2.1: A function $f : X \to \mathbb{R}$, identification of critical points and critical values, and the corresponding persistence diagram (corresponding to the sublevel sets filtration).

2.2.2 Computing degree-0 persistence

Input: Graph (V, E), and a map $f : V \sqcup E \to \mathbb{R}$.

Hypothesis: f gives us a *graph filtration*, that is for any $(u, v) \in E$ where $u, v \in V$, we have $f((u, v)) \geq \max\{f(u), f(v)\}$.

Pre-processing: Sort $V \sqcup E$ by increasing lexicographic order (value of f, dimension). It gives a sequence $(\sigma_1, \ldots, \sigma_m)$ of vertices and edges. Then initialize a union-find data structure $\mathcal{V}$.

```
for $i = 1 \ldots m$ do
    if $\sigma_i$ is a vertex $v$ then
        Create new entry $e_v := \{v\}$ in $\mathcal{V}$                // birth of a new cc;
    else                                                          // ($\sigma_i$ is an edge $(u, v)$)
        Find entries $e_u, e_v$ containing respectively $u$ and $v$ in $\mathcal{V}$;
        if $e_u \neq e_v$ then                                     // assume wlog $f(e_u) < f(e_v)$
            Merge $e_v$ into $e_u$ in $\mathcal{V}$;
            Record the interval $[f(e_v), f((u, v)))$ in the barcode ;
        end
    end
end
```

Algorithm 1: Degree-0 persistence

Post-processing: For any e_v remaining in $\mathcal{V}$, record $[f(e_v), +\infty)$.

Running time:

- Pre-processing : $\mathcal{O}(m \log(m))$
- Main: $\mathcal{O}(m\alpha(m))$, where α is the inverse Ackerman function.
- Post-processing: $\mathcal{O}(m)$.