# VI - Statistical aspects of persistent homology

*PSL Week - Topological Data Analysis*

**Abstract**

We discuss how persistent homology behaves under random sampling. We highlight a notion of low intrinsic dimension called the $(a, b)$-standard assumption, and show how to leverage the stability of persistence, as well as elementary minimax theory, to study the problem of estimation of persistence diagrams.

## Contents



## 1 Sampling assumptions and $(a, b)$-standard measures

### 1.1 Distance function and Hausdorff distance

Let $K \subset \mathbb{R}^d$ be compact. We recall the *distance function* to $K$.

**Definition 1.1** (Distance function)**.** The distance to $K$ is

$$\mathrm{dist}(\cdot,) : \mathbb{R}^d \to [0, \infty), \qquad \mathrm{dist}(x, K) := \min_{p \in K} \|x - p\|.$$

**Definition 1.2** (Offset)**.** For $r > 0$, the *r-offset* (or *thickening*) of $K$ is

$$K^r := \{x \in \mathbb{R}^d : \mathrm{dist}(x, K) \leq r\} = \bigcup_{p \in K} \overline{B}_r(p).$$

**Definition 1.3** (Hausdorff distance)**.** Let $A, B$ be compact subsets of $\mathbb{R}^d$. The *Hausdorff distance* between $A$ and $B$ is

$$d_H(A, B) := \min\{r \geq 0 : A \subset B^r \text{ and } B \subset A^r\}.$$

One can show the equivalent expression

$$d_H(A, B) = \sup_{x \in \mathbb{R}^d} |\mathrm{dist}(x, A) - \mathrm{dist}(x, B)|.$$

Intuitively, $d_H(A, B)$ is the smallest radius so that every point of $A$ is within $r$ of $B$ and every point of $B$ is within $r$ of $A$ (see Figure 1).

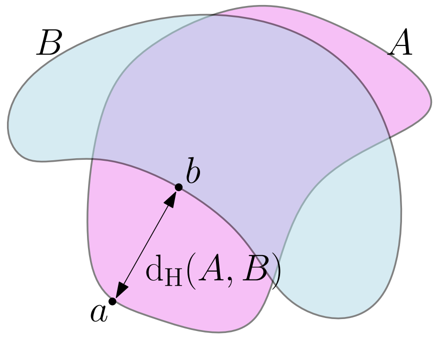


Figure 1: The Hausdorff distance between two subsets $A$ and $B$ of the plane. In this example, $d_H(A, B)$ is the distance between the point $a$ in $A$ which is the farthest from $B$ and its nearest neighbor $b$ on $B$.

### 1.2 $(a, b)$-standard measures

We now put a regularity condition on the sampling distribution.

**Definition 1.4** ($(a, b)$-standard measure)**.** Let $P$ be a Borel probability measure on $\mathbb{R}^d$. We say that $P$ is $(a, b)$-*standard at scale* $r_0$ if there exist constants $a > 0$, $b > 0$, $r_0 > 0$ such that for all $x \in \text{support}(P)$ and all $0 < r \leq r_0$,

$$P\big(B(x, r)\big) \;\geq\; a\, r^b.$$

*Remark* 1.5. Roughly speaking, a measure that is $(a, b)$-standard behaves on small scales like the $b$-dimensional Lebesgue measure:

- for $b = d$ and $P$ having a density bounded below on its support, this condition is satisfied;
- $b$ does not need to be an integer: this covers fractal-like supports (e.g. Cantor-type sets).

The exponent $b$ plays the role of an *effective dimension*: balls of radius $r$ carry at least a constant times $r^b$ mass, so the support cannot be too thin.

To quantify how "massive" the support is, it is convenient to introduce covering and packing numbers.

**Definition 1.6** (Covering and packing numbers)**.** Let $K \subset \mathbb{R}^d$ be bounded and $r > 0$.

- An $r$-*covering* of $K$ is a family of balls of radius $r$ whose union contains $K$. The *covering number* is
$$\text{cov}(K, r) := \min\Big\{k : K \subset \bigcup_{i=1}^{k} B(x_i, r)\Big\}.$$
- An $r$-*packing* of $K$ is a family of disjoint balls of radius $r$ with centres in $K$. The *packing number* is
$$\text{pack}(K, r) := \max\big\{k : \exists x_1, \ldots, x_k \in K,\; B(x_i, r) \text{ disjoint}\big\}.$$

**Proposition 1.7** (Massiveness of $(a, b)$-standard measures)**.** *Let $P$ be $(a, b)$-standard at scale $r_0$ and $K = \text{support}(P)$. Then there exists a constant $C_{a,b} > 0$ such that for all $r \leq r_0$,*

$$\text{pack}(K, r) \;\leq\; \frac{1}{ar^b}, \qquad \text{cov}(K, r) \;\leq\; \frac{C_{a,b}}{r^b}.$$

*Idea.* If $B(x_1, r), \ldots, B(x_N, r)$ is a maximal packing, then these balls are disjoint and all contained in $K^r$, so

$$1 = P(\mathbb{R}^d) \ \geq \ \sum_{i=1}^{N} P(B(x_i, r)) \ \geq \ N\, ar^b,$$

which yields $N \leq 1/(ar^b)$. Using duality between coverings and packings (i.e. $\mathrm{pack}(K, r) \leq \mathrm{cov}(K, r) \leq \mathrm{pack}(K, r/2)$) gives the covering bound. □

Thus a $(a, b)$-standard measure has at most on the order of $r^{-b}$ well-separated points at scale $r$, just like a $b$-dimensional cube $[0, 1]^b$.

## 2 Hausdorff convergence of samples and consequences

### 2.1 A non-asymptotic bound

Let $P$ be $(a, b)$-standard at scale $r_0$, and let $X_1, \ldots, X_n$ be i.i.d. with distribution $P$. We denote the sample by

$$\mathcal{X}_n := \{X_1, \ldots, X_n\} \subset \mathbb{R}^d.$$

The following proposition shows that $\mathcal{X}_n$ converges to the support of $P$ in Hausdorff distance, with a rate governed by $b$.

**Proposition 2.1** (Hausdorff convergence under $(a, b)$-standardness)**.** *Let $P$ be $(a, b)$-standard at scale $r_0$, with compact support $K = \mathrm{support}(P)$. Let $\mathcal{X}_n$ be an i.i.d. sample from $P$. Then:*

*(a) There exist constants $C_{a,b,\alpha} > 0$ such that for any $\alpha > 0$ and all $n$ large enough,*

$$\mathbb{P}\Big(d_H(K, \mathcal{X}_n) \ > \ \big(C_{a,b,\alpha} \tfrac{\log n}{n}\big)^{1/b}\Big) \ \leq \ n^{-\alpha}.$$

*(b) Equivalently, for any confidence level $\delta \in (0, 1)$ and any $r \leq 2r_0$, one has*

$$\mathbb{P}\big(d_H(K, \mathcal{X}_n) \leq r\big) \ \geq \ 1 - \delta$$

*as soon as*

$$n \ \geq \ \frac{C'_{a,b}}{r^b}\Big(\log \tfrac{1}{r} + \log \tfrac{1}{\delta}\Big).$$

*Idea.* Fix $r \leq 2r_0$ and consider a minimal $(r/2)$-covering of $K$ with $N = \mathrm{cov}(K, r/2) \lesssim r^{-b}$ balls $B_1, \ldots, B_N$ of radius $r/2$. If $d_H(K, \mathcal{X}_n) > r$, one easily checks that at least one ball $B_j$ contains no sample point. Since $P(B_j) \geq ar^b$, the probability that $B_j$ is empty is at most $(1 - a(r/2)^b)^n \leq \exp(-an(r/2)^b)$. A union bound over all $j$ then gives

$$\mathbb{P}\big(d_H(K, \mathcal{X}_n) > r\big) \ \leq \ N \exp(-an(r/2)^b) \ \lesssim \ r^{-b} \exp(-an(r/2)^b).$$

Optimizing in $r$ yields the rate $r_n \asymp (\log n/n)^{1/b}$ and the stated bounds. □

In words: for an $(a, b)$-standard measure, with $n$ points we typically resolve the support down to a scale of order $(\log n/n)^{1/b}$ in Hausdorff distance.

## 2.2 Plug-in estimation of persistence diagrams

We now want to transfer the Hausdorff convergence of $\mathcal{X}_n$ to convergence of persistence diagrams. Let $(M, \rho)$ be a compact metric space, and let Filt be a filtration functor that associates to each compact subset $A \subset M$ a filtration of simplicial complexes $\mathrm{Filt}(A)$ (e.g. the Vietoris–Rips or Čech filtration). Under mild assumptions, recall from Chapter III that persistent homology is *stable* with respect to perturbations of $A$ in the Hausdorff metric.

**Theorem 2.2** (Stability for spaces (informal)). *Let $(M, \rho)$ be a compact metric space and $A, B \subset M$ compact. For a fixed homological degree $k$, let $D_k(A)$ and $D_k(B)$ denote the persistence diagrams of $H_k(\mathrm{Filt}(A))$ and $H_k(\mathrm{Filt}(B))$ (with coefficients in a field). Then,*

$$d_B\big(D_k(A), D_k(B)\big) \;\le\; d_H(A, B),$$

*where $d_B$ is the bottleneck distance.*

We now combine Proposition 2.1 and Theorem 2.2. Let $(M, \rho)$ be a compact metric space and let $\mu$ be a Borel probability measure on $M$ with compact support $X_\mu := \mathrm{support}(\mu) \subset M$. Let $X_1, \ldots, X_n$ be i.i.d. with distribution $\mu$ and $\mathcal{X}_n := \{X_1, \ldots, X_n\}$.

**Definition 2.3** (Statistical model). Fix $a, b > 0$. We denote by $\mathcal{P}_{M,a,b}$ the collection of Borel probability measures $\mu$ on $M$ such that:

- the support $X_\mu$ is compact in $M$;
- $\mu$ is $(a, b)$-standard (with respect to $\rho$).

**Theorem 2.4** (Upper bounds for persistence diagrams). *Assume $\mu \in \mathcal{P}_{M,a,b}$ and let $D_k(\mu)$ denote the $k$-th persistence diagram of $\mathrm{Filt}(X_\mu)$, and $D_k(\mathcal{X}_n)$ that of $\mathrm{Filt}(\mathcal{X}_n)$. Then:*

*(a) For all $\varepsilon > 0$,*

$$\mathbb{P}\big(d_B(D_k(\mu), D_k(\mathcal{X}_n)) > C\,\varepsilon\big) \;\le\; \min\Big(\tfrac{C'}{\varepsilon^b}\exp(-cn\varepsilon^b),\; 1\Big),$$

*where $C, C', c > 0$ depend only on the filtration and on $a, b$.*

*(b) For $n$ large enough,*

$$\sup_{\mu \in \mathcal{P}_{M,a,b}} \mathbb{E}_{\mu^n}\left[d_B(D_k(\mu), D_k(\mathcal{X}_n))\right] \;\le\; C_{a,b}\left(\frac{\log n}{n}\right)^{1/b},$$

*where $C_{a,b}$ depends only on $a, b$ and the filtration.*

*Idea.* For each $\mu$, by stability for spaces,

$$d_B(D_k(\mu), D_k(\mathcal{X}_n)) \;\le\; C\, d_H(X_\mu, \mathcal{X}_n).$$

Apply Proposition 2.1 with $K = X_\mu$, then take the supremum over $\mu \in \mathcal{P}_{M,a,b}$ and integrate the tail bound to control the expectation by using $\mathbb{E}[Y] = \int_0^\infty \mathbb{P}(Y > y)\mathrm{d}y$ for a random variable $Y \ge 0$. □

Thus the usual estimator $D_k(\mathcal{X}_n)$ is consistent, and its accuracy improves at the rate $(\log n/n)^{1/b}$, up to constants.

# 3 Minimax risk and Le Cam's lemma

As standard in statistical decision theory, we now turn to the question of optimality: Does there exist any better estimator than mine? Said otherwise: How well can *any* estimator do, in the worst case, over a given statistical model?

### 3.1 Risk and minimax risk

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space (the observation space). A *statistical model* is a collection $\mathcal{Q}$ of probability measures on $(\mathcal{X}, \mathcal{A})$.

We are interested in a parameter

$$\theta : \mathcal{Q} \to \Theta,$$

where $(\Theta, \rho)$ is a metric space (for us, $\Theta$ will be a space of persistence diagrams, and $\rho$ the bottleneck distance).

An *estimator* is a measurable map

$$\hat{\theta}_n : \mathcal{X}^n \to \Theta,$$

applied to $n$ i.i.d. observations $X_1, \ldots, X_n \sim Q$.

**Definition 3.1** (Risk and minimax risk)**.** The *risk* of $\hat{\theta}_n$ at $Q$ (for loss $\rho$) is

$$R(Q, \hat{\theta}_n) := \mathbb{E}_{Q^n}\big[\rho\big(\theta(Q), \hat{\theta}_n(X_1, \ldots, X_n)\big)\big].$$

The *minimax risk* over $\mathcal{Q}$ is

$$R_n(\mathcal{Q}) := \inf_{\hat{\theta}_n} \sup_{Q \in \mathcal{Q}} R(Q, \hat{\theta}_n),$$

where the infimum is over all estimators $\hat{\theta}_n : \mathcal{X}^n \to \Theta$.

$R_n(\mathcal{Q})$ measures the best possible worst-case performance for the estimation problem $(\mathcal{Q}, \theta)$ at sample size $n$.

### 3.2 Total variation distance

We will use the *total variation* distance between probability measures.

**Definition 3.2** (Total variation)**.** Let $Q, Q'$ be probability measures on $(\mathcal{X}, \mathcal{A})$. The *total variation* distance is

$$\mathrm{TV}(Q, Q') := \sup_{A \in \mathcal{A}} |Q(A) - Q'(A)|.$$

If $Q$ and $Q'$ admit densities $q, q'$ with respect to a reference measure $\nu$, then

$$\mathrm{TV}(Q, Q') = \frac{1}{2} \int_{\mathcal{X}} |q - q'| \, \mathrm{d}\nu.$$

*Remark* 3.3. For product measures, one can show

$$\mathrm{TV}(Q^n, Q'^n) \ \le\ 1 - (1 - \mathrm{TV}(Q, Q'))^n.$$

In particular, if $\mathrm{TV}(Q, Q')$ is small, then $\mathrm{TV}(Q^n, Q'^n)$ remains bounded away from 1 for moderate $n$.

### 3.3 Le Cam's two-point lemma

Le Cam's lemma is a simple but powerful way to get minimax lower bounds, by restricting attention to just two distributions $Q, Q' \in \mathcal{Q}$.

**Lemma 3.4** (Le Cam)**.** *Let $\mathcal{Q}$ be a set of probability measures on $(\mathcal{X}, \mathcal{A})$, and $\theta : \mathcal{Q} \to \Theta$ a parameter, where $(\Theta, \rho)$ is a metric space. Then for any $Q, Q' \in \mathcal{Q}$,*

$$R_n(\mathcal{Q}) = \inf_{\hat{\theta}_n} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n} \rho(\theta(Q), \hat{\theta}_n) \ \ge\ \frac{1}{2}\, \rho\big(\theta(Q), \theta(Q')\big) \, \big(1 - \mathrm{TV}(Q, Q')\big)^n.$$

*Proof.* Fix any estimator $\hat{\theta}_n$, and any $Q, Q' \in \mathcal{Q}$. The key observation is to notice that

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n} \rho(\theta(Q), \hat{\theta}_n) \;\geq\; \frac{1}{2} \left( \mathbb{E}_{Q^n} \rho(\theta(Q), \hat{\theta}_n) + \mathbb{E}_{Q'^n} \rho(\theta(Q'), \hat{\theta}_n) \right).$$

Then, changing the measure under which these two integral are taken and using the triangle inequality,

$$\rho(\theta(Q), \hat{\theta}_n) + \rho(\hat{\theta}_n, \theta(Q')) \geq \rho(\theta(Q), \theta(Q')).$$

This means that the estimator cannot simultaneously do very well under $Q^n$ and under $Q'^n$ if these two distributions are difficult to distinguish statistically. After that, we obtain

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n} \rho(\theta(Q), \hat{\theta}_n) \;\geq\; \frac{1}{2} \rho(\theta(Q), \theta(Q')) \left(1 - \mathrm{TV}(Q^n, Q'^n)\right),$$

and the proof is completed using a data-processing inequality for testing. □

Informally: if two distributions in the model have parameters that are far apart, but are close in total variation, then any estimator must incur a nontrivial error on at least one of them.

# 4 Minimax rates for persistence

We now apply this framework to persistence diagrams of random point clouds sampled from $(a, b)$-standard measures.

## 4.1 Setup and upper bound

We consider the model $\mathcal{P}_{M,a,b}$ defined above, with parameter of interest

$$\theta(\mu) := D_k(\mu),$$

the $k$-th persistence diagram of $\mathrm{Filt}(X_\mu)$, and loss

$$\rho(D, D') := d_B(D, D').$$

A natural estimator is

$$\widehat{\theta}_n := D_k(\mathcal{X}_n),$$

the diagram built on the empirical point cloud.

Theorem 2.4 shows that this estimator satisfies

$$\sup_{\mu \in \mathcal{P}_{M,a,b}} \mathbb{E}_{\mu^n} d_B\big(D_k(\mu), D_k(\mathcal{X}_n)\big) \;\leq\; C_{a,b} \left( \frac{\log n}{n} \right)^{1/b}.$$

In particular,

$$R_n(\mathcal{P}_{M,a,b}) \;\leq\; C_{a,b} \left( \frac{\log n}{n} \right)^{1/b},$$

so the minimax risk cannot be worse than this rate.

### 4.2 Lower bound via Le Cam

We now sketch a lower bound, following a two-point argument. Assume that $(M, \rho)$ is a metric space in which we can find a point $x \in M$ and a sequence $(x_n)_{n\geq 1} \subset M$ such that

$$\rho(x, x_n) \asymp (an)^{-1/b}.$$

(This holds in particular in $\mathbb{R}^d$ with $b \leq d$, by choosing a grid of points with spacing of order $n^{-1/b}$.)

For each $n$, consider the two measures

$$\mu_0 = \delta_x, \qquad \mu_{1,n} := \Big(1 - \frac{1}{n}\Big)\delta_x + \frac{1}{n}\,\delta_{x_n}.$$

One can check that both $\mu_0$ and $\mu_{1,n}$ belong to $\mathcal{P}_{M,a,b}$ for suitable constants $a, b$ (they are $(a, b)$-standard, since balls around $x$ or $x_n$ quickly contain all mass).

Let $D_k(\mu_0)$ and $D_k(\mu_{1,n})$ be the corresponding persistence diagrams (for a fixed filtration and degree $k$). Geometrically, $\mu_0$ has a single-point support $\{x\}$, while $\mu_{1,n}$ has two points $\{x, x_n\}$ at distance $\rho(x, x_n) \asymp n^{-1/b}$.

- The bottleneck distance between $D_k(\mu_0)$ and $D_k(\mu_{1,n})$ is of order $\rho(x, x_n)$: the presence of the extra point $x_n$ creates additional small features in the filtration at scale $\rho(x, x_n)$, so that

$$d_B\big(D_k(\mu_0), D_k(\mu_{1,n})\big) \gtrsim \rho(x, x_n) \asymp n^{-1/b}.$$

- The total variation distance between $\mu_0$ and $\mu_{1,n}$ is exactly

$$\mathrm{TV}(\mu_0, \mu_{1,n}) = \frac{1}{n},$$

so

$$\big(1 - \mathrm{TV}(\mu_0, \mu_{1,n})\big)^n = \Big(1 - \frac{1}{n}\Big)^n \longrightarrow e^{-1}.$$

Applying Le Cam's lemma

$$R_n(\mathcal{P}_{M,a,b}) \geq \frac{1}{2}\, d_B\big(D_k(\mu_0), D_k(\mu_{1,n})\big)\, \big(1 - \mathrm{TV}(\mu_0, \mu_{1,n})\big)^n$$

and using the two bullets above, we obtain

$$R_n(\mathcal{P}_{M,a,b}) \gtrsim n^{-1/b}.$$

**Theorem 4.1** (Minimax lower bound for persistence)**.** *Under the assumptions above, there exists a constant $c > 0$ such that for all $n$ large enough,*

$$R_n(\mathcal{P}_{M,a,b}) \geq c\, n^{-1/b}.$$

Combining this with the upper bound from Theorem 2.4, we obtain that the estimator $D_k(\mathcal{X}_n)$ is *minimax optimal up to logarithmic factors*:

$$c\, n^{-1/b} \lesssim R_n(\mathcal{P}_{M,a,b}) \leq C_{a,b} \left(\frac{\log n}{n}\right)^{1/b}.$$

*Remark* 4.2. The log factor in the upper bound comes from the union bound over an $r$-covering of the support and is typical in nonparametric estimation with $(a, b)$-standard assumptions. For $b = 1$, it cannot actually be removed.