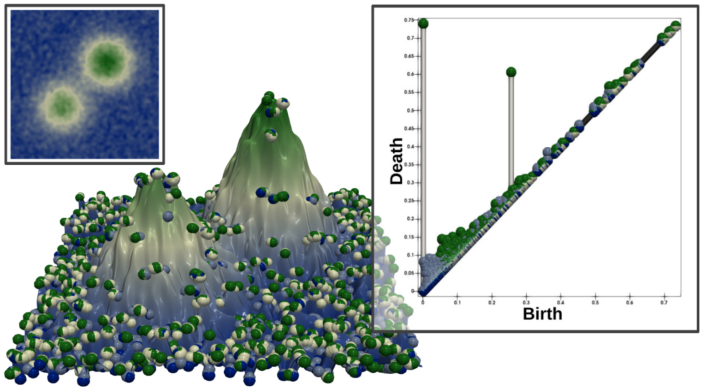
Mémoire de 1ère année

Département de Mathématiques et Applications
École Normale Supérieure, Paris

Analyse topologique des données

Clara BRIAND
Léon GUILHOT

Sous la direction de
Eddie AAMARI & Emmanuel GIROUX



Juin 2025

Table des matières

I Persistance **2**

1 Homologie simpliciale **2**

2 Modules de persistance **4**

3 Diagramme d'un module de persistance **6**

4 Persistance topologique **8**

5 Stabilité des diagrammes de persistance **10**

II Estimation statistique **16**

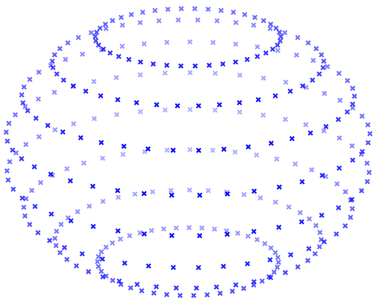
6 Filtrations classiques **16**

7 Résultats probabilistes **17**

8 Estimation du diagramme de persistance **20**

Introduction

Considérons l'image suivante.



On peut y voir huit anneaux parallèles, ou bien une sphère, ou bien même 342 points dans l'espace. Tout dépend de l'échelle à laquelle on regarde l'image. Lorsqu'on analyse des données, le choix de la bonne échelle pour les étudier est fondamental. Or, en grande dimension, on ne peut plus représenter ces points dans l'espace. On a donc besoin d'un autre moyen de visualiser leur structure. C'est précisément pour répondre à ce défi que *l'analyse topologique des données* s'est développée au début des années 2000. Cette discipline s'appuie sur la topologie pour extraire et comprendre la forme intrinsèque des données, même lorsqu'elles sont complexes et de grande dimension. Plutôt que de se limiter à des mesures classiques, elle met en lumière des caractéristiques globales telles que des composantes connexes, des cycles ou des cavités. Un concept central de cette théorie est celui de la *persistance*, introduit et formalisé principalement par Herbert Edelsbrunner, Gunnar Carlsson et Jean Harer. La persistance permet de suivre la naissance et la disparition des caractéristiques topologiques des données à différentes échelles. Les *diagrammes de persistance* résument ces informations en un objet plus facile à manipuler, tout en reflétant fidèlement la structure topologique des données qu'ils représentent. C'est précisément cet objet que nous allons étudier dans ce mémoire.

Dans la première partie, on définira quelques notions d'algèbre et de topologie algébrique essentielles à la compréhension de notre mémoire. On commencera notamment par parler d'homologie simpliciale, puis on introduira à l'aide de la théorie des carquois le concept de module de persistance qui nous permettra ensuite de définir les diagrammes de persistance associés à ces modules. On définira la *distance d'entrelacement*, notée d_i, entre les modules de persistance et la *distance goulot de bouteille*, notée d_b, entre les diagrammes de persistance. Le cœur de cette partie sera de démontrer le théorème suivant qui nous garantit la stabilité de ces diagrammes.

Théorème (d'isométrie) *Soient $\mathbb{V}, \mathbb{W}$ des modules de persistance q-modérés sur $\mathbb{R}$, alors*

$$d_b\big(\mathrm{dgm}(\mathbb{V}), \mathrm{dgm}(\mathbb{W})\big) = d_i(\mathbb{V}, \mathbb{W}).$$

Dans la deuxième partie on s'intéressera à la question de l'estimation statistique des diagrammes de persistance. Plus précisément à partir d'un échantillonnage d'un compact $K \subset \mathbb{R}^d$, on construira un estimateur du diagramme de persistance d'une filtration associée à K. On démontrera une borne sur sa vitesse de convergence et on se posera également la question de son optimalité.

Notre mémoire s'appuie en grande partie sur le livre *Persistence Theory : From Quiver Representations to Data Analysis* de Steve Oudot [4], ainsi que sur les articles [1] et [2] de Frédéric Chazal, Vin de Silva, Marc Glisse et Steve Oudot, et également sur les notes du cours *Geometric Inference* donné par Eddie Aamari en 2024. On en profite pour le remercier, ainsi qu'Emmanuel Giroux pour l'encadrement de ce mémoire.

Première partie
Persistance

1 Homologie simpliciale

Définition 1.1. Un *k-simplexe* σ est l'enveloppe convexe de $k+1$ points $p_0, ..., p_k \in \mathbb{R}^d$ affinement indépendants, c'est-à-dire qu'ils ne sont contenus dans aucun sous-espace de dimension inférieure ou égale à $k-1$. Les points $p_0, ..., p_k$ sont appellés les *sommets* du simplexe σ. On notera dans la suite $\sigma = [p_0, ..., p_k]$ le k-simplexe défini par les points $p_0, ..., p_k \in \mathbb{R}^d$.

De manière équivalente, le k-simplexe $\sigma = [p_0, ..., p_k]$ est l'ensemble des combinaisons linéaires $\sum_{i=0}^{k} \lambda_i p_i$ avec $\lambda_i \in \mathbb{R}$ et $\sum_{i=0}^{k} \lambda_i = 1$. Une *$\ell$-face* d'un k-simplexe $\sigma = [p_0, ..., p_k]$ est un ℓ-simplexe $\tau = [p_{i_0}, ..., p_{i_\ell}]$. Ainsi un 0-simplexe est un point, un 1-simplexe est un segment, un 2-simplexe est un triangle et un 3-simplexe est un tétraèdre. Les 1-faces d'un 2-simplexe $\sigma = [p_0, p_1, p_2]$ sont les arêtes $[p_0, p_1]$, $[p_0, p_2]$ et $[p_1, p_2]$.

Définition 1.2. Un *complexe simplicial* $\mathcal{K}$ de $\mathbb{R}^d$ est une collection de simplexes qui satisfait les propriétés suivantes :
— chaque face d'un simplexe de $\mathcal{K}$ est un simplexe de $\mathcal{K}$.
— l'intersection de deux simplexes de $\mathcal{K}$ est soit vide, soit une face commune.

Un complexe simplicial est dit fini s'il est constitué d'un nombre fini de simplexes. Dans la suite, les complexes simpliciaux avec lesquels on travaillera seront toujours finis.

Définition 1.3. La *dimension* d'un complexe simplicial est la dimension maximale de ses simplexes.

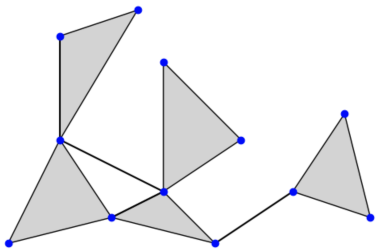


FIGURE 1.1 – Un complexe simplicial constitué de cinq 2-simplexes.

Soit $\mathcal{K}$ un complexe simplicial. Une *k-chaîne* de $\mathcal{K}$ est une somme formelle $\sum_{i=0}^{r} \varepsilon_i \sigma_i$ où les $\sigma_0, ..., \sigma_r$ sont les k-simplexes de $\mathcal{K}$ et les ε_i sont à valeurs dans $\mathbb{Z}/2\mathbb{Z}$. Soient $\lambda \in \mathbb{Z}/2\mathbb{Z}$, $c = \sum_{i=0}^{r} \varepsilon_i \sigma_i$, $c' = \sum_{i=0}^{r} \varepsilon'_i \sigma_i$, on peut définir la somme de deux k-chaînes et le produit par un scalaire de la manière suivante : $c + c' = \sum_{i=0}^{r} (\varepsilon_i + \varepsilon'_i)\sigma_i$, et $\lambda \cdot c = \sum_{i=0}^{r} \lambda \varepsilon_i \sigma_i$ où la somme et le produit sont pris modulo 2. La k-chaîne nulle est $\sum_{i=0}^{r} 0 \cdot \sigma_i$. On note ainsi $\mathcal{C}_k(\mathcal{K})$ l'espace vectoriel des k-chaînes de $\mathcal{K}$. Une base de cet espace vectoriel est donnée par l'ensemble des k-simplexes de $\mathcal{K}$.

Définition 1.4. Le *bord* d'un k-simplexe σ est la somme de ses $(k-1)$-faces (qui est une $(k-1)$-chaîne si on voit σ comme un complexe simplicial). On note le bord de $\sigma = [p_0, ..., p_k]$

$$\partial_k(\sigma) = \sum_{i=0}^{k} [p_0, ..., \hat{p_i}, ..., p_k],$$

où $[p_0, ..., \hat{p_i}, ..., p_k]$ désigne la $(k-1)$-face définie par les sommets $p_0, ..., p_k$ à l'exception de p_i.

On peut étendre cette application linéairement et ainsi définir l'opérateur bord

$$\partial_k : \left\{ \begin{array}{rcl} \mathcal{C}_k(\mathcal{K}) & \longrightarrow & \mathcal{C}_{k-1}(\mathcal{K}) \\ c & \longmapsto & \sum_{\sigma \in c} \partial(\sigma) \end{array} \right. .$$

Proposition 1.5. $\partial_{k-1} \circ \partial_k = 0$.

Preuve. L'opérateur bord étant linéaire, on vérifie l'égalité pour un simplexe $\sigma = [p_0, ..., p_k]$.

$$\begin{aligned} \partial_{k-1} \circ \partial_k(\sigma) &= \partial_{k-1}\Big(\sum_{i=0}^{k} [p_0, ..., \hat{p_i}, ..., p_k] \Big) = \sum_{i=0}^{k} \partial_{k-1}\big([p_0, ..., \hat{p_i}, ..., p_k] \big) \\ &= \sum_{0 \leq j < i \leq k} [p_0, ..., \hat{p_j}, ..., \hat{p_i}, ..., p_k] + \sum_{0 \leq i < j \leq k} [p_0, ..., \hat{p_i}, ..., \hat{p_j}, ..., p_k] \\ &= 2 \sum_{0 \leq j < i \leq k} [p_0, ..., \hat{p_j}, ..., \hat{p_i}, ..., p_k] = 0. \end{aligned}$$

□

Définition 1.6. Le *complexe de chaînes* associé à un complexe simplicial $\mathcal{K}$ de dimension d est la suite

$$0 \longrightarrow \mathcal{C}_d(\mathcal{K}) \xrightarrow{\partial_d} ... \xrightarrow{\partial_{k+1}} \mathcal{C}_k(\mathcal{K}) \xrightarrow{\partial_k} ... \xrightarrow{\partial_1} \mathcal{C}_0(\mathcal{K}) \longrightarrow 0$$

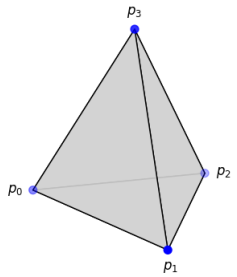
Pour tout entier $k \in [\![0, d]\!]$ on appelle *ensemble des k-cycles de* $\mathcal{K}$, l'ensemble $\mathcal{Z}_k(\mathcal{K}) = \ker \partial_k$ et *ensemble des k-bords de* $\mathcal{K}$, l'ensemble $\mathcal{B}_k(\mathcal{K}) = \operatorname{im} \partial_{k+1}$. Ces ensembles sont des sous-espaces vectoriels de $\mathcal{C}_k(\mathcal{K})$ et la proposition précédente nous dit que $\mathcal{B}_k(\mathcal{K}) \subseteq \mathcal{Z}_k(\mathcal{K})$. On n'a en général pas l'égalité et donc le complexe de chaînes associé à $\mathcal{K}$ n'est pas toujours une suite exacte. L'homologie permet justement de mesurer ce défaut d'exactitude.

Définition 1.7. On appelle *k-ième groupe d'homologie de* $\mathcal{K}$ (ou *groupe d'homologie de dimension k* de $\mathcal{K}$), l'espace vectoriel

$$\mathcal{H}_k(\mathcal{K}) = \mathcal{Z}_k(\mathcal{K}) / \mathcal{B}_k(\mathcal{K}).$$

Sa dimension, notée $\beta_k(\mathcal{K})$, est appelée *k-ième nombre de Betti.*

Exemple 1.8. Calculons les trois premiers nombres de Betti du complexe simplicial ci-dessous.



Le complexe de chaînes associé à ce complexe simplicial est le suivant :

$$0 \xrightarrow{\partial_3} \mathcal{C}_2 \xrightarrow{\partial_2} \mathcal{C}_1 \xrightarrow{\partial_1} \mathcal{C}_0 \xrightarrow{\partial_0} 0$$

$\mathcal{C}_0$ a pour base l'ensemble des quatre points, $\mathcal{C}_1$ a pour base l'ensemble des six arêtes et $\mathcal{C}_2$ a pour base l'ensemble des quatre faces. Les sommets sont ordonnés comme sur le dessin ci-dessus et on ordonne les arêtes et les faces de la manière suivante :

$$a_0 = [p_0, p_1] \;\; a_1 = [p_0, p_2] \;\; a_2 = [p_0, p_3] \;\; a_3 = [p_1, p_2] \;\; a_4 = [p_1, p_3] \;\; a_5 = [p_2, p_3],$$

$$\sigma_0 = [p_0, p_1, p_2] \;\; \sigma_1 = [p_0, p_1, p_3] \;\; \sigma_2 = [p_0, p_2, p_3] \;\; \sigma_3 = [p_1, p_2, p_3].$$

Les morphismes ∂_1 et ∂_2 sont donc définis par les matrices

$$\partial_1 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix} \quad \text{et} \quad \partial_2 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

On a donc $\operatorname{rang} \partial_1 = 3$ et $\operatorname{rang} \partial_2 = 3$, et le théorème du rang nous dit que $\dim \ker \partial_1 = 6 - 3 = 3$ et $\dim \ker \partial_2 = 4 - 3 = 1$. On a ainsi $\beta_0 = 4 - 3 = 1$, $\beta_1 = 3 - 3 = 0$ et $\beta_2 = 1 - 0 = 1$.

On a introduit dans une définition précédente la notion de complexe de chaînes associé à un complexe simplicial. Un complexe de chaînes d'espaces vectoriels sur $\mathbb{Z}/2\mathbb{Z}$ est en fait un objet purement algébrique défini comme étant une suite de $\mathbb{Z}/2\mathbb{Z}-$espaces vectoriels indexée sur $\mathbb{Z}$

$$\ldots \xrightarrow{\partial_{n+2}} \mathcal{C}_{n+1} \xrightarrow{\partial_{n+1}} \mathcal{C}_n \xrightarrow{\partial_n} \mathcal{C}_{n-1} \xrightarrow{\partial_{n-1}} \ldots$$

où chaque application ∂_n est linéaire et $\partial_n \circ \partial_{n+1} = 0$. Étant donné deux complexes de chaînes $\mathcal{C}, \mathcal{C}'$, un morphisme $f : (\mathcal{C}, \partial) \to (\mathcal{C}', \partial')$ de complexes de chaînes est la donnée pour chaque $n \in \mathbb{Z}$ d'une application linéaire $f_n : \mathcal{C}_n \to \mathcal{C}'_n$ tel que le diagramme suivant commute.

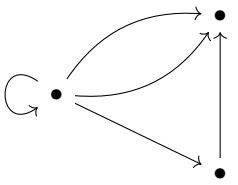
$$\begin{array}{ccccccccc} \ldots & \xrightarrow{\partial_{n+2}} & \mathcal{C}_{n+1} & \xrightarrow{\partial_{n+1}} & \mathcal{C}_n & \xrightarrow{\partial_n} & \mathcal{C}_{n-1} & \xrightarrow{\partial_{n-1}} & \ldots \\ & & \downarrow{\scriptstyle f_{n+1}} & & \downarrow{\scriptstyle f_n} & & \downarrow{\scriptstyle f_{n-1}} & & \\ \ldots & \xrightarrow{\partial'_{n+2}} & \mathcal{C}'_{n+1} & \xrightarrow{\partial'_{n+1}} & \mathcal{C}'_n & \xrightarrow{\partial'_n} & \mathcal{C}'_{n-1} & \xrightarrow{\partial'_{n-1}} & \ldots \end{array}$$

On a donc défini la catégorie des complexes de chaînes d'espaces vectoriels sur $\mathbb{Z}/2\mathbb{Z}$ que l'on note $\mathbf{Ch}_{\mathbb{Z}/2\mathbb{Z}}$. On peut maintenant étendre la définition de l'homologie simpliciale donnée plus haut en un foncteur $\mathcal{H}_n : \mathbf{Ch}_{\mathbb{Z}/2\mathbb{Z}} \to \mathbf{Vect}_{\mathbb{Z}/2\mathbb{Z}}$ qui à un complexe de chaînes $(\mathcal{C}, \partial)$ associe l'espace vectoriel $\ker \partial_n / \operatorname{im} \partial_{n+1}$, et à un morphisme $f : (\mathcal{C}, \partial) \to (\mathcal{C}', \partial')$ associe l'application linéaire induite $\mathcal{H}_n(f) : \mathcal{H}_n\big((\mathcal{C}, \partial)\big) \to \mathcal{H}_n\big((\mathcal{C}', \partial')\big)$. On peut en effet vérifier que $\mathcal{H}_n$ préserve les compositions et les identités. Ainsi si $f : \mathcal{K} \to \mathcal{K}'$ est une application simpliciale (c'est-à-dire qu'elle envoie un simplexe de $\mathcal{K}$ sur un simplexe de $\mathcal{K}'$), alors f induit naturellement un morphisme entre les complexes de chaînes associés et induit donc un morphisme $f^k_* : \mathcal{H}_k(\mathcal{K}) \to \mathcal{H}_k(\mathcal{K}')$.

2 Modules de persistance

Les modules de persistance sont un certain type de diagramme d'espaces vectoriels et d'applications linéaires. On les introduit en utilisant le langage de la théorie des carquois.

Définition 2.1. Un *carquois* est un multi-graphe orienté avec possiblement une infinité de sommets et d'arêtes.



Définition 2.2. Un *carquois linéaire de type* L_n est un graphe linéaire orienté à n sommets dans lequel toutes les flèches sont orientées de la gauche vers la droite.

$$\bullet \longrightarrow \bullet \longrightarrow \ldots \longrightarrow \bullet \longrightarrow \bullet$$

Définition 2.3. Une *représentation* $\mathbb{V}$ d'un carquois Q sur un corps k consiste à associer à chaque sommet i du carquois Q un k-espace vectoriel V_i et à chaque flèche reliant un sommet i à un sommet j une application k-linéaire $v_\alpha : V_i \to V_j$.

Définition 2.4. Un *morphisme de représentation* $\phi : \mathbb{V} \to \mathbb{W}$ entre deux représentations d'un carquois Q sur un corps k est une collection d'applications k-linéaires $\phi_i : V_i \to W_i$ telles que pour tous sommets i et j et toute flèche α reliant les sommets i et j, le diagramme suivant commute.

$$\begin{array}{ccc} V_i & \xrightarrow{v_\alpha} & V_j \\ {\scriptstyle \phi_i}\downarrow & & \downarrow{\scriptstyle \phi_j} \\ W_i & \xrightarrow[w_\alpha]{} & W_j \end{array}$$

Le morphisme ϕ est un *isomorphisme de représentation* si chacune des applications linéaires ϕ_i est un isomorphisme d'espace vectoriel.

Définition 2.5. La *somme directe* de deux représentations $\mathbb{V}$ et $\mathbb{W}$ d'un carquois Q sur un corps k est la représentation $\mathbb{V} \oplus \mathbb{W}$ où les espaces vectoriels sont les $V_i \oplus W_i$ et les applications k-linéaires sont les $v_\alpha \oplus w_\alpha$. La représentation $\mathbb{V}$ est dite *indécomposable* si elle ne peut pas s'écrire comme somme directe de représentations non triviales.

On peut également définir les représentations pour un sous-poset du poset $(\mathbb{R}, \leq)$. Soit $T \subseteq \mathbb{R}$, en voyant le poset $(T, \leq)$ comme une catégorie où les objets sont les éléments de T et les morphismes caractérisent la relation d'ordre, on peut définir une représentation de ce poset comme un foncteur de cette catégorie vers la catégorie des espaces vectoriels. Une représentation de $(T, \leq)$ est donc un ensemble d'espaces vectoriels V_i où $i \in T$ et un ensemble d'applications k-linéaires $v_i^j : V_i \to W_j$ où $i \leq j$, tel que pour tout $i \in T$, $v_i^i = id_{V_i}$ et pour tous $i \leq j \leq k$, $v_i^k = v_j^k \circ v_i^j$.

Définition 2.6. Une représentation $\mathbb{V}$ d'un carquois Q est dite de *dimension finie* si Q est fini et si la dimension de $\mathbb{V}$ (définie comme étant la somme des dimensions des espaces vectoriels V_i qui la composent) est finie. Elle est dite de *dimension finie en chaque point* lorsque chacun des espaces vectoriels V_i est de dimension finie.

Pour un carquois de type L_n, on définit la *représentation d'intervalle* $\mathbb{I}_Q[b, d]$ sur un corps k, par la représentation indécomposable décrite ci-dessous où l'indice b désigne la première apparition de k et l'indice d la dernière apparition de k. On parlera parfois simplement *d'intervalle* $\mathbb{I}_Q[b, d]$.

$$0 \xrightarrow{0} \ldots \xrightarrow{0} 0 \xrightarrow{0} k \xrightarrow{id_k} \ldots \xrightarrow{id_k} k \xrightarrow{0} 0 \xrightarrow{0} \ldots \xrightarrow{0} 0$$

Ces représentations sont particulièrement utiles puisqu'elles vont nous permettre sous certaines hypothèses de décomposer de manière simple des représentations plus compliquées. Le théorème suivant a été prouvé dans les années 1970 par Krull et Gabriel. Une preuve plus moderne, due à Henry Hale, peut être trouvée dans [6].

Théorème 2.7. (Krull - Gabriel) *Toute représentation de dimension finie* $\mathbb{V}$ *d'un carquois* Q *de type* L_n *sur un corps* k*, se décompose de manière unique en une somme directe de représentations d'intervalles* $\mathbb{I}_Q[b_i, d_i]$.

Une représentation d'intervalle $\mathbb{I}[b, d]$ sur un corps k du sous-poset $(T, \leq)$ du poset $(\mathbb{R}, \leq)$ est définie de la manière suivante : $V_i = k$ pour tout $i \in [b, d]$ et $V_i = 0$ sinon. De même $v_i^j = id_k$ pour tous $i, j \in [b, d]$ et $v_i^j = 0$ sinon. On définit de façon analogue les représentations $\mathbb{I}]b, d]$,

$\mathbb{I}[b,d[$ et $\mathbb{I}]b,d[$. On peut ainsi étendre le théorème précédent aux représentations de dimension finie en chaque point du sous-poset $(T,\leq)$. La preuve du théorème suivant peut être trouvée dans [3].

Théorème 2.8. (Crawley - Boevey) *Toute représentation sur un corps k de dimension finie en chaque point d'un sous-poset $(T,\leq)$ du poset $(\mathbb{R},\leq)$, se décompose de manière unique en une somme directe de représentations d'intervalles.*

On est maintenant prêts à introduire la notion de module de persistance.

Définition 2.9. Soient $T \subseteq \mathbb{R}$, k un corps, un *module de persistance* sur T est une représentation du poset $(T,\leq)$.

Ces modules de persistance peuvent, sous certaines hypothèses, être représentés par des diagrammes que l'on appelle *diagrammes de persistance*.

3 Diagramme d'un module de persistance

Lorsqu'un module de persistance se décompose de façon unique en somme directe d'intervalles (par exemple lorsqu'il est de dimension finie en chaque point), on peut le représenter à l'aide d'un *diagramme de persistance* où chaque point de coordonnée (b,d) sur le diagramme désigne l'intervalle correspondant dans la décomposition du module en somme directe de représentations d'intervalles. Les intervalles qui caractérisent ces représentations peuvent être ouverts, fermés, semi-ouverts à droite ou à gauche et avec des bornes possiblement infinies. Les points du diagramme peuvent donc être décorés pour indiquer la forme de l'intervalle selon la règle suivante :

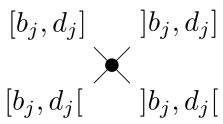


FIGURE 3.1 – Règle de décoration des diagrammes - extrait de Oudot [4].

Lorsqu'il existe, le diagramme de persistance d'une représentation $\mathbb{V}$ est noté $\mathrm{Dgm}(\mathbb{V})$, la version non décorée est quant à elle notée $\mathrm{dgm}(\mathbb{V})$. On notera parfois les intervalles de la manière suivante :

$$[b,d] = (b^-,d^+)\ ;\ [b,d[= (b^-,d^-)\ ;\ ]b,d] = (b^+,d^+)\ ;\ ]b,d[= (b^+,d^-).$$

Exemple 3.1. On considère le module de persistance $\mathbb{V}$ qui admet la décomposition en somme directe de représentations d'intervalles suivante :

$$\mathbb{V} = \mathbb{I}[1,4] \oplus \mathbb{I}]2,3] \oplus \mathbb{I}[3,5[\oplus \mathbb{I}[3,4].$$

Le diagramme de persistance associé à cette décomposition est donc le suivant :

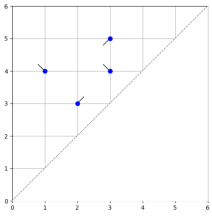


FIGURE 3.2 – Diagramme de persistance associé à la représentation $\mathbb{V}$.

On va maintenant définir une mesure sur ces diagrammes. Soit $\mathbb{V}$ un module de persistance décomposable en somme directe de représentations d'intervalles, et soit $\mathrm{Dgm}(\mathbb{V})$ son diagramme de persistance associé. Soit $R = [p, q] \times [r, s]$ un rectangle avec $-\infty \leq p < q \leq r < s \leq +\infty$, on dit qu'un point $(b^{\pm}, d^{\pm})$ de $\mathrm{Dgm}(\mathbb{V})$ appartient au rectangle R si $[q, r] \subseteq (b^{\pm}, d^{\pm}) \subseteq [p, s]$. On note alors $\mathrm{Dgm}(\mathbb{V})|_R$ l'ensemble des points de $\mathrm{Dgm}(\mathbb{V})$ appartenant au rectangle R. Un point situé sur le bord d'un rectangle appartient à ce rectangle si sa décoration pointe vers l'interieur.

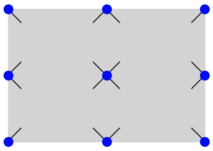


FIGURE 3.3 – Illustration de la relation d'appartenance.

La *mesure de comptage* de ce rectangle est définie par $\tilde{\mu}_{\mathbb{V}}(R) = \#\,\mathrm{Dgm}(\mathbb{V})|_R$.

On remarque que la mesure de comptage est additive : si un rectangle R s'écrit comme union, disjointe sauf sur un bord, de deux rectangles R_1 et R_2, alors $\tilde{\mu}_{\mathbb{V}}(R) = \tilde{\mu}_{\mathbb{V}}(R_1) + \tilde{\mu}_{\mathbb{V}}(R_2)$. Les rectangles formant une base de voisinages du demi-plan $\pi^+ = \{(x, y) \in \overline{\mathbb{R}}^2 \mid y > x\}$, le théorème d'extension de Kolmogorov nous garantit l'existence d'une unique mesure sur π^+ qui prolonge $\tilde{\mu}$.

On peut utiliser cette idée pour définir une mesure sur π^+ qui nous permettra de construire les diagrammes de persistance pour des modules qui ne sont pas décomposables en intervalles.

Définition 3.2. Lorsque $\mathbb{V} = (V_i, v_i^j)$ est un module de persistance indexé sur $\mathbb{R}$, on dit que $\mathbb{V}$ est *q-modéré* si le rang de chacun des v_i^j est fini.

Définition 3.3. Soient $\mathbb{V}$ un module de persistance q-modéré, et $-\infty < p < q \leq r < s < +\infty$. On définit la *mesure du rang* d'un rectangle de la manière suivante :

$$\begin{aligned}
\mu_{\mathbb{V}}\big([-\infty, q] \times [r, +\infty]\big) &= rg(v_q^r), \\
\mu_{\mathbb{V}}\big([-\infty, q] \times [r, s]\big) &= rg(v_q^r) - rg(v_q^s), \\
\mu_{\mathbb{V}}\big([p, q] \times [r, +\infty]\big) &= rg(v_q^r) - rg(v_p^r), \\
\mu_{\mathbb{V}}\big([p, q] \times [r, s]\big) &= rg(v_q^r) - rg(v_q^s) + rg(v_p^s) - rg(v_p^r).
\end{aligned}$$

Notons que la condition de q-modération nous garantit qu'il n'y a pas d'indéterminations du type $\infty - \infty$.

Proposition 3.4. *Soit $\mathbb{V}$ un module de persistance q-modéré et décomposable en intervalles, soit $R \subseteq \pi^+$ un rectangle, alors $\tilde{\mu}_{\mathbb{V}}(R) = \mu_{\mathbb{V}}(R)$.*

Preuve. On considère d'abord un rectangle R de la forme $[-\infty, q] \times [r, +\infty]$, avec $q \leq r$. Un point (x, y) de $\mathrm{Dgm}(\mathbb{V})$ sera dans R si et seulement si $x \geq q$ et $y \leq r$. Il correspond à une représentation d'intervalle $\mathbb{I}[x, y]$ avec $[x, y] \subseteq [q, r]$. L'application v_q^r est obtenue comme somme directe des applications qui interviennent entre les espaces en position q et r dans la décomposition en intervalles de $\mathbb{V}$. Or ceci est exactement le nombre d'intervalles $\mathbb{I}[x, y]$ avec $[x, y] \subseteq [q, r]$ qui interviennent dans cette décomposition. On a donc bien

$$\tilde{\mu}_{\mathbb{V}}\big([-\infty, q] \times [r, +\infty]\big) = rg(v_q^r).$$

Les trois autres égalités se déduisent de la même façon en utilisant le principe d'inclusion-exclusion. □

Comme la mesure de comptage est positive et additive, On en déduit que la mesure du rang l'est aussi, et on peut alors la prolonger en une mesure sur π^+. Cela nous permet de définir les diagrammes de persistance pour des modules indexés sur $\mathbb{R}$ qui ne sont pas décomposables en somme directe de représentations d'intervalles. C'est l'objet du théorème ci-dessous dont la preuve peut être trouvée dans la section 2.4 de [1].

Théorème 3.5. *Soit $\mathbb{V}$ un module de persistance q-modéré. Il existe un unique ensemble localement fini de points décorés de $\pi^+ = \{(x, y) \in \overline{\mathbb{R}}^2 \mid x < y\}$, que l'on note $\mathrm{Dgm}(\mu_{\mathbb{V}})$, tel que pour tout rectangle $R = [p, q] \times [r, s]$ avec $-\infty \leq p < q < r < s \leq +\infty$, on ait $\mu_{\mathbb{V}}(R) = \#\,\mathrm{Dgm}(\mu_{\mathbb{V}})|_R$.*

Ce qu'il faut retenir de cette section est que pour un module de persistance $\mathbb{V}$ indexé sur T, son diagramme de persistance associé est bien défini dans les trois situations suivantes :

- $\mid T \mid < \infty$
- $\mathbb{V}$ est de dimension finie en chaque point
- $T = \mathbb{R}$ et $\mathbb{V}$ est q-modéré

De plus, si un module de persistance $\mathbb{V}$ est à la fois q-modéré et décomposable en somme directe de représentations d'intervalles, alors le diagramme obtenu par le théorème précédent et celui obtenu par sa décomposition en somme directe de représentations d'intervalles coïncident, sauf éventuellement sur la diagonale.

4 Persistance topologique

On a défini dans la première section l'homologie pour les complexes simpliciaux. On peut aussi définir l'homologie pour des espaces topologiques quelconques, on appelle cela l'homologie singulière. On note Δ_k le k-simplexe standard dans $\mathbb{R}^k$ (c'est-à-dire le k-simplexe défini par les vecteurs de la base canonique de $\mathbb{R}^k$) et on appelle *k-simplexe singulier* d'un espace topologique X une application continue $\sigma : \Delta_k \to X$. L'homologie singulière de X se construit alors d'une façon similaire à l'homologie simpliciale. On ne détaillera pas plus cette construction mais on notera simplement qu'il est possible de parler d'homologie pour des espaces topologiques quelconques. Cette définition est fonctorielle, c'est-à-dire que toute application continue $f : X \to Y$ induit un morphisme $f_*^k : \mathcal{H}_k(X) \to \mathcal{H}_k(Y)$. Si de plus X et Y sont homotopiquement équivalents, le morphisme induit est alors un isomorphisme.

Définition 4.1. Soit X un espace topologique, soit $\mathcal{X} = \{X_i \mid i \in \mathbb{R}\} \subseteq \mathcal{P}(X)$ tel que $i \leq j$ implique $X_i \subseteq X_j$. On dit que $\mathcal{X}$ est une *filtration* de X si $\cap_{i\in\mathbb{R}} X_i = \emptyset$ et $\cup_{i\in\mathbb{R}} X_i = X$. On dit alors que $(X, \mathcal{X})$ est un *espace filtré*.

Si $(X, \mathcal{X})$ est un espace filtré sur $T \subseteq \mathbb{R}$, on dispose d'une suite $(X_i)_{i\in T}$ d'espaces topologiques inclus les uns dans les autres, où l'on note $v_i^j : X_i \to X_j$ les inclusions pour tous $i < j$. On peut alors appliquer le foncteur d'homologie de dimension p pour obtenir un module de persistance que l'on note $\mathcal{H}_p(\mathcal{X})$. La collection $\mathcal{H}_*(\mathcal{X}) = \{\mathcal{H}_p(\mathcal{X}) \mid p \geq 0\}$ est appelée *homologie persistante* de la filtration $\mathcal{X}$. Si ces modules de persistance sont q-modérés ou bien de dimension finie en chaque point, on peut leur associer leur diagramme de persistance. Dans ce cas-là, la filtration est dite q-modérée ou de dimension finie en chaque point. La superposition de ces diagrammes est appelée *diagramme de persistance de la filtration* $\mathcal{X}$ et est noté $\mathrm{dgm}(\mathcal{X})$.

Exemple 4.2. On prend pour espace topologique X le graphe de la fonction f représenté ci-dessous. On va définir sur ce graphe une filtration appelée *filtration par sous-niveaux de la fonction f*. Pour tout $i \in \mathbb{R}$, on note $X_i = f^{-1}\big(]-\infty, i]\big)$, l'ensemble $\mathcal{X} = \{X_i \mid i \in \mathbb{R}\}$ définit alors une filtration sur l'espace topologique X. En effet on a d'une part $\cap_{i\in\mathbb{R}} X_i = f^{-1}(\emptyset) = \emptyset$ et d'autre part $\cup_{i\in\mathbb{R}} X_i = f^{-1}\big(]-\infty, +\infty[\big) = X$.

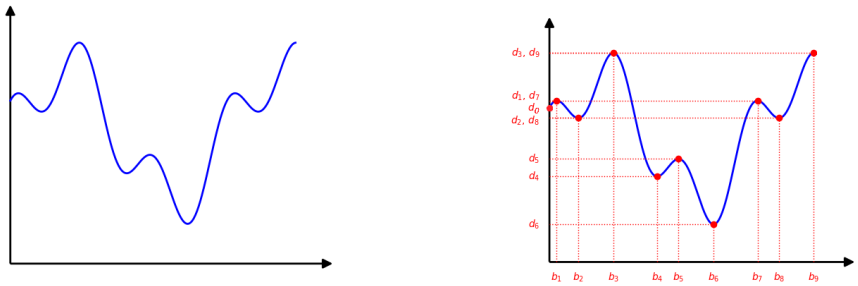


FIGURE 4.1 – Graphe de la fonction f et coordonnées de ses extrema.

Nous n'avons pas particulièrement parlé d'homologie pour des espaces topologiques quelconques, notons simplement que les espaces X_i étant des sous-ensembles de $\mathbb{R}$, leur homologie de dimension $n \geq 1$ est triviale. On ne s'intéresse donc qu'à l'homologie de dimension 0, pour laquelle les nombres de Betti $\beta_0(X_i)$ correspondent au nombre de composantes connexes des X_i. Par application du foncteur d'homologie, on obtient un module de persistance indexé sur $\mathbb{R}$ où les espaces vectoriels sont les $\mathcal{H}_0(X_i)$ et les morphismes $\mathcal{H}_0(v_i^j)$ sont les morphismes induits par les inclusions. Le nombre de composantes connexes des X_i augmente aux minima locaux et diminue aux maxima locaux de la fonction f. Ce module de persistance est donc de dimension finie en chaque point et par conséquent le théorème 2.8 nous assure qu'il est décomposable de manière unique en somme directe de représentations d'intervalle.

Il nous faut maintenant déterminer ces intervalles, dont les bornes sont données par les valeurs des minima et maxima de la fonction f. La figure 4.1 nous donne justement ces valeurs de minima et maxima. Lorsqu'une nouvelle composante connexe apparaît (à un minimum local d_i), on ouvre un intervalle de persistance. Lorsqu'une fusion de composantes a lieu (à un maximum local d_j), on ferme l'un des intervalles en cours. Plus précisément, on applique la règle usuelle en homologie persistante dite du « plus jeune meurt en premier » [1] : lorsqu'une composante disparaît par fusion, c'est celle qui est apparue le plus récemment qui « meurt ». On ferme donc l'intervalle correspondant à la composante la plus jeune. On a une première composante connexe qui apparaît à l'indice d_6, puis une autre à l'indice d_4. On ouvre donc deux intervalles à d_6 et d_4. Ces deux composantes connexes se rejoignent à l'indice d_5, on doit donc fermer l'intervalle commencé à d_4 en vertu de la règle énoncée plus haut. On obtient alors notre premier intervalle $[d_4, d_5[$. On peut continuer ainsi de suite et on obtient finalement la décomposition

$$\mathcal{H}_0(\mathcal{X}) = \mathbb{I}[d_6, +\infty[\oplus \mathbb{I}[d_4, d_5[\oplus \mathbb{I}[d_2, d_3[\oplus \mathbb{I}[d_8, d_7[\oplus \mathbb{I}[d_0, d_1[.$$

Ce qui nous donne le diagramme de persistance ci-dessous.

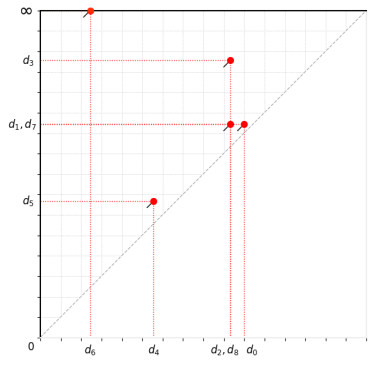


FIGURE 4.2 – Diagramme de persistance de la filtration par sous-niveaux de la fonction f.

1. Cette règle, dite *Elder Rule*, n'est bien sûr pas arbitraire en raison du théorème 2.8. Cependant, il n'en existe pas de démonstration précise dans la littérature. J.Curry en propose tout de même une dans un cas particulier [5].

5 Stabilité des diagrammes de persistance

Le théorème que l'on va démontrer dans cette section est le cœur de la théorie de la persistance. Il nous garantit que les objets que l'on a construits précédemment ont les propriétés de stabilité espérées, c'est-à-dire, qu'à deux modules de persistance « proches » correspondent des diagrammes de persistance « ressemblants », et réciproquement.

5.1 Métriques

Dans cette partie, on va formaliser ce que veut dire que deux modules ou diagrammes de persistance se ressemblent. Intuitivement, deux diagrammes de persistance sont proches s'il suffit de faire des petits déplacements pour passer d'un point à l'autre. C'est ce que quantifie la distance *goulot de bouteille*, que l'on introduit ci-dessous.

Définition 5.1. Soient A et B des multi-ensembles de points de $\overline{\mathbb{R}}^2$. Une *correspondance partielle* de (A, B) est un sous-ensemble $M \subseteq A \times B$ tel que :

1. Pour tout $a \in A$, il existe au plus un $b \in B$ tel que $(a, b) \in M$,
2. Pour tout $b \in B$, il existe au plus un $a \in A$ tel que $(a, b) \in M$.

Par convention, s'il existe un élément apparaissant, par exemple, plusieurs fois dans le multi-ensemble A, chacune de ses occurrences peut être mise en correspondance avec un élément de B. On notera M_{sc} le multi-ensemble des points de $A \sqcup B$ qui n'ont pas de correspondant. Étant donnée une correspondance partielle $M \subseteq A \times B$, on définit le coût d'une paire $(a, b) \in M$ par $c(a, b) = \|a-b\|_\infty$, et le coût d'un point sans correspondant $d = (d_x, d_y) \in M_{sc}$ par $c(d) = \frac{1}{2}|d_x - d_y|$, c'est-à-dire la distance en norme infinie entre ce point et la droite d'équation $x - y = 0$. On définit alors le coût de la correspondance M par :

$$c(M) = \max\Big\{ \sup_{(a,b)\in M} c(a, b), \sup_{d\in M_{sc}} c(d)\Big\}.$$

Définition 5.2. Soient P et Q des diagrammes de persistance non décorés. La *distance goulot de bouteille* entre P et Q est notée $d_b(P, Q)$ et est définie par

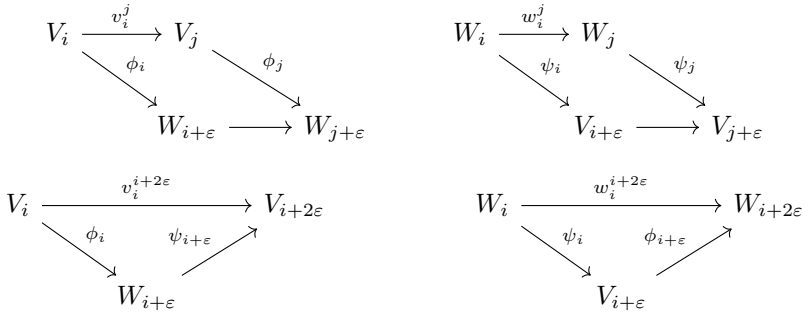
$$d_b(P, Q) = \inf\big\{c(M) \mid M \text{ est une correspondance partielle de } (P, Q)\big\}.$$

Proposition 5.3. *La distance goulot de bouteille satisfait l'inégalité triangulaire.*

Preuve. Soient P, R, S des diagrammes de persistance et M_1, M_2 des correspondances partielles de (P, R) et (R, S) respectivement. On peut construire une correspondance partielle M de (P, S) par : $(a, c) \in M$ si et seulement s'il existe $b \in R$ tel que $(a, b) \in M_1$ et $(b, c) \in M_2$. On peut vérifier au cas par cas que $c(M) \leq c(M_1) + c(M_2)$. □

On définit maintenant la *distance d'entrelacement* entre deux modules de persistance. Celle-ci nous indique à quel point ces modules sont loin d'être isomorphes.

Définition 5.4. Soit $\varepsilon > 0$. On dit que deux modules de persistance $\mathbb{V}$ et $\mathbb{W}$ sont *ε-entrelacés* si pour tout $t \in \mathbb{R}$, il existe des morphismes $\phi_t : V_t \longrightarrow W_{t+\varepsilon}$ et $\psi_t : W_t \longrightarrow V_{t+\varepsilon}$ tels que les diagrammes suivants commutent :



Définition 5.5. On définit la *distance d'entrelacement*, notée d_i, sur l'ensemble des modules de persistance par

$$d_i(\mathbb{V}, \mathbb{W}) = \inf \big\{ \varepsilon > 0 \mid \mathbb{V}, \mathbb{W} \text{ sont } \varepsilon\text{-entrelacés} \big\}.$$

La distance d'entrelacement satisfait l'inégalité triangulaire (on peut en trouver une preuve dans la section 4.1 de [1]), mais n'est cependant pas une vraie distance. En effet, les représentations d'intervalles $\mathbb{I}]0,1[$ et $\mathbb{I}[0,1]$ sont à distance 0, mais ne sont même pas isomorphes.

5.2 Le théorème d'isométrie

Dans cette section, on va énoncer et démontrer le théorème d'isométrie.

Théorème 5.6. (d'isométrie) *Soient $\mathbb{V}$ et $\mathbb{W}$ des modules de persistance q-modérés sur $\mathbb{R}$.*

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

Notons que la condition de q-modération nous garantit, par le théorème 3.5, que les diagrammes de persistance sont bien définis.

La preuve est longue, et par moments assez technique, mais permet d'introduire beaucoup d'idées récurrentes dans la théorie de la persistance. On la divise en plusieurs lemmes, dont on donne les idées générales de preuve sans vérifier tous les détails, lesquels peuvent être consultés dans [1].

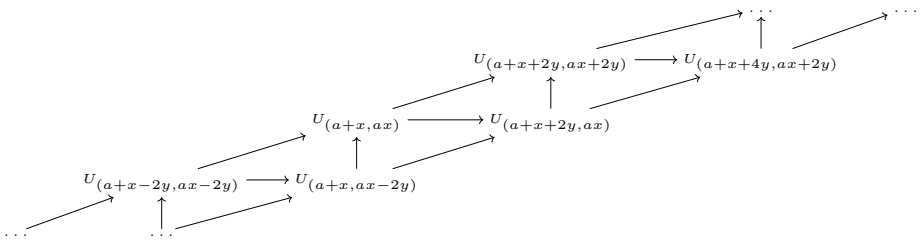
5.2.1 Inégalité de stabilité

Pour la suite, on se fixe $\mathbb{V}$ et $\mathbb{W}$ comme dans l'énoncé du théorème, et on commence par prouver l'inégalité dite de *stabilité* $d_b(\mathrm{dgm}\ \mathbb{V}, \mathrm{dgm}\ \mathbb{W}) \leq d_i(\mathbb{V}, \mathbb{W})$.

Le résultat suivant joue à lui seul un rôle important dans la théorie de la persistance.

Lemme 5.7. (d'interpolation) *Soit $\varepsilon > 0$, si $\mathbb{V}$ et $\mathbb{W}$ sont ε-entrelacés, alors il existe une famille $(\mathbb{U}_x)_{x\in[0,\varepsilon]}$ de modules de persistance tels que $\mathbb{U}_0 = \mathbb{V}$, $\mathbb{U}_\varepsilon = \mathbb{W}$ et pour tous $x, y \in [0,\varepsilon]$ les modules $\mathbb{U}_x$ et $\mathbb{U}_y$ sont $|x-y|$-entrelacés.*

Preuve. Pour $r \in \mathbb{R}$, on définit la *diagonale translatée* Δ_r comme étant la droite d'équation $x - y = 2r$, et on munit le plan $\overline{\mathbb{R}}^2$ d'un ordre partiel $\preceq$ en posant pour tous $(x,y), (x'y') \in \overline{\mathbb{R}}^2$, $(x,y) \preceq (x',y')$ si $x \leq x'$ et $y \leq y'$. De cette façon, on obtient une identification canonique entre les modules de persistance sur $\mathbb{R}$ et les représentations du poset $(\Delta_r, \preceq)$, donnée par l'isomorphisme (de représentations) $t \in \mathbb{R} \mapsto (t-r, t+r) \in \Delta_r$. On fixe maintenant $x, y \in [0,\varepsilon]$. On remarque que le problème de trouver des représentations $\mathbb{U}_x$ et $\mathbb{U}_y$ $|x-y|$-entrelacées se réduit à trouver une représentation $\mathbb{U}$ de $(\Delta_x \cap \Delta_y, \preceq)$ telle que $\mathbb{U}_{|\Delta_x} \cong \mathbb{U}_x$ et $\mathbb{U}_{|\Delta_y} \cong \mathbb{U}_y$. En effet, l'ordre partiel qu'on a défini sur le plan nous garantit que le diagramme suivant commute pour tout $a \in \mathbb{R}$.



Ceci se traduit, avec l'isomorphisme construit plus haut, à des représentations $|x-y|$-entrelacées. Notre objectif est donc de construire une représentation $\mathbb{U}$ sur la bande $\Delta_{[0,\varepsilon]} = \{p \in \Delta_r \mid r \in [0,\varepsilon]\}$ consistante avec l'ordre $\preceq$, et satisfaisant $\mathbb{U}_{|\Delta_x} \cong \mathbb{U}_x$ et $\mathbb{U}_{|\Delta_y} \cong \mathbb{U}_y$. La famille recherchée sera

donnée par les représentations sur $\mathbb{R}$ identifiées à la restriction de $\mathbb{U}$ à chaque diagonale translatée. On définit sur $(\mathbb{R}^2, \preceq)$ les représentations suivantes :

- $\mathbb{A}$ définie par $A_{(x,y)} = V_x$
- $\mathbb{B}$ définie par $B_{(x,y)} = W_{y-\varepsilon}$
- $\mathbb{C}$ définie par $C_{(x,y)} = V_y$
- $\mathbb{D}$ définie par $D_{(x,y)} = W_{x+\varepsilon}$

où les morphismes sont induits par ceux de $\mathbb{V}$ et $\mathbb{W}$. On construit également les morphismes entre ces représentations, où ϕ_t et ψ_t sont les morphismes donnés par l'ε-entrelacement entre $\mathbb{V}$ et $\mathbb{W}$:

- $\mathrm{Id}_{\mathbb{V}} : \mathbb{A} \longrightarrow \mathbb{C}$, défini en (x, y) par $v_x^y : V_x \longrightarrow V_y$
- $\Phi : \mathbb{A} \longrightarrow \mathbb{D}$, défini en (x, y) par $\phi_x : V_x \longrightarrow W_{x+\varepsilon}$
- $\Psi : \mathbb{B} \longrightarrow \mathbb{C}$, défini en (x, y) par $\psi_{y-\varepsilon} : W_{y-\varepsilon} \longrightarrow V_y$
- $\mathrm{Id}_{\mathbb{W}} : \mathbb{B} \longrightarrow \mathbb{D}$, défini en (x, y) par $w_{x+\varepsilon}^{y-\varepsilon} : W_{y-\varepsilon} \longrightarrow W_{x+\varepsilon}$

Ces morphismes sont tous bien définis sur $\Delta_{[0,\varepsilon]}$, et ils permettent de construire un nouveau morphisme $M : \mathbb{A} \oplus \mathbb{B} \longrightarrow \mathbb{C} \oplus \mathbb{D}$, donné par la matrice

$$\begin{bmatrix} \mathrm{Id}_{\mathbb{V}} & \Psi \\ \Phi & \mathrm{Id}_{\mathbb{W}} \end{bmatrix}.$$

On définit finalement $\mathbb{U}$ comme étant la représentation induite par M, qui à chaque point $(x, y) \in \Delta_{[0,\varepsilon]}$, associe l'espace vectoriel image $\mathrm{im}_M(\mathbb{A} \oplus \mathbb{B})_{(x,y)}$, avec les morphismes correspondants.

Pour montrer que $\mathbb{U}$ est la représentation que l'on cherchait, il faut vérifier que $\mathbb{U}_{|\Delta_x} \cong \mathbb{V}$ et $\mathbb{U}_{|\Delta_\varepsilon} \cong \mathbb{W}$. Le détail de ce calcul peut être consulté dans la section 3.4 de [1].

□

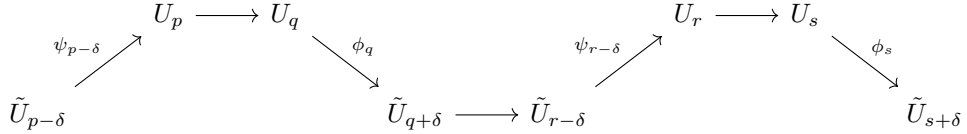
Pour la suite, on fixe $\varepsilon > 0$ et on suppose que $\mathbb{V}$ et $\mathbb{W}$ sont ε-entrelacés. On note $(\mathbb{U}_x)_{x\in[0,\varepsilon]}$ la famille de représentations qui interpole $\mathbb{V}$ et $\mathbb{W}$, dont l'existence est garantie par le lemme d'interpolation.

L'étape suivante de la preuve est de relier les distances d'entrelacement et de goulot de bouteille, qui sont assez peu intuitives, à d'autres quantités que l'on connaît mieux. On commence donc par exprimer la distance d'entrelacement à l'aide de la mesure du rang μ, définie dans la section 2.

Lemme 5.8. (Inégalités des boîtes) *Soient $\delta > 0$ et $\mathbb{U}$, $\tilde{\mathbb{U}}$ deux modules de persistance δ-entrelacées. Soient $p, q, r, s \in \overline{\mathbb{R}}$ tels que $r > q + 2\delta$. On définit le rectangle $R = [p, q] \times [r, s] \subseteq \overline{\mathbb{R}}^2$ et son δ-épaississement $R^\delta = [p - \delta, q + \delta] \times [r - \delta, s + \delta] \subseteq \overline{\mathbb{R}}^2$. On a alors*

$$\mu_{\mathbb{U}}(R) \leq \mu_{\tilde{\mathbb{U}}}(R^\delta),$$
$$\mu_{\tilde{\mathbb{U}}}(R) \leq \mu_{\mathbb{U}}(R^\delta).$$

Preuve. On montre ici le cas où p, q, r et s sont finis, les autres cas se prouvent avec la même idée. Par hypothèse, $q + \delta < r - \delta$, donc le diagramme suivant est bien défini :



où pour $t \in \mathbb{R}$, $\phi_t : \mathbb{U}_t \longrightarrow \tilde{\mathbb{U}}_{t+\delta}$ et $\psi_t : \tilde{\mathbb{U}}_t \longrightarrow \mathbb{U}_{t+\delta}$ sont les morphismes fournis par le δ-entrelacement entre $\mathbb{U}$ et $\tilde{\mathbb{U}}$. Par définition de la mesure du rang, on a :

$$\begin{aligned} \mu_{\tilde{U}}(R^\delta) &= rg(\tilde{u}_{r-\delta}^{q+\delta}) + rg(\tilde{u}_{s+\delta}^{p-\delta}) - rg(\tilde{u}_{r-\delta}^{p-\delta}) - rg(\tilde{u}_{s+\delta}^{q+\delta}) \\ &\leq rg(u_r^q) + rg(u_s^p) - rg(u_r^p) - rg(u_s^q) = \mu_{\mathbb{U}}(R). \end{aligned}$$

□

On relie maintenant la distance goulot de bouteille à la pseudo-distance de Hausdorff.

Définition 5.9. Soient P un diagramme de persistance et $x \in \overline{\mathbb{R}}^2$. La distance d entre P et x est donnée par

$$d(x,P) = \min\{\|x - \pi_{\text{diag}}(x)\|_2, \sup_{y\in P} \|x-y\|_2\}$$

où π_{diag} est la projection sur la droite d'équation $x = y$. C'est le minimum entre la distance usuelle d'un point à un fermé, et la distance euclidienne de x à la diagonale.

Définition 5.10. Soient P, Q des diagrammes de persistance. La *pseudo-distance de Hausdorff* entre P et Q, notée $d_H(P,Q)$, est

$$d_H(P,Q) = \max\{\sup_{x\in Q} d(x,P), \sup_{y\in P} d(y,Q)\}.$$

On parle ici de pseudo-distance car P et Q sont des multi-ensembles de points non bornés et contenant même l'infini.

Lemme 5.11. *Soient $x, y \in \mathbb{R}$ tels que $0 \leq x \leq y \leq \varepsilon$. On suppose que* dgm $\mathbb{U}_x$ *et* dgm $\mathbb{U}_y$ *ont un nombre fini de points. Alors*

$$d_H(\text{dgm } \mathbb{U}_x, \text{dgm } \mathbb{U}_y) \leq |x-y|.$$

De plus, il existe $\delta_0 > 0$ tel que si $|x-y| < \delta_0$, alors

$$d_b(\text{dgm } \mathbb{U}_x, \text{dgm } \mathbb{U}_y) \leq |x-y|.$$

Preuve. On pose $\delta = |x-y|$ et $(a,b) \in$ dgm $\mathbb{U}_x$. Pour tout $k > 0$, on définit le carré $C_k = [a-k, a+k] \times [b-k, b+k]$ de côté $2k$, centré en (a,b). Comme $(a,b) \in C_k$, on a $\mu_{\mathbb{U}_x}(C_k) > 0$. Supposons d'abord que pour k assez petit, C_k^δ ne rencontre pas la diagonale. Par les inégalités des boîtes, $\mu_{\mathbb{U}_y}(C_k^\delta) > 0$, il y a donc au moins un point de dgm $\mathbb{U}_y$ dans le δ-épaississement de C_k. En faisant tendre k vers 0, on conclut que $d\big((a,b), \text{dgm } \mathbb{U}_y\big) \leq \delta$.

Si C_k^δ rencontre la diagonale pour tout $k > 0$, alors la distance euclidienne entre celle-ci et (a,b) est plus petite que δ. Par définition de la pseudo-distance de Hausdorff, on aura aussi dans ce cas que $d\big((a,b), \text{dgm } \mathbb{U}_y\big) \leq \delta$. En inversant les rôles de $\mathbb{U}_x$ et $\mathbb{U}_y$, on a également que $d\big((a,b), \text{dgm } \mathbb{U}_y\big) \leq \delta$. Par conséquent, $d_H(\text{dgm } \mathbb{U}_x, \text{dgm } \mathbb{U}_y) \leq \delta$, ce qui prouve la première inégalité.

Montrons maintenant la deuxième inégalité. La distance minimale entre deux points distincts de dgm $\mathbb{U}_x$ est bien définie, puisque par hypothèse dgm $\mathbb{U}_x$ est fini, on la note ξ_1. La distance minimale entre un point de dgm $\mathbb{U}_x$ et la diagonale est bien définie aussi, on la note ξ_2. Soit $\delta_0 < \frac{1}{3}\min\{\delta_1, \delta_2\}$, et $|x-y| = \delta < \delta_0$. On prend $0 < \xi < \delta$ un paramètre que l'on fera tendre vers 0 plus tard. Alors $\mu_{\mathbb{U}_x}(C_\xi) = \mu_{\mathbb{U}_y}(C_\xi^{2\delta})$ et il y a donc le même nombre de points de dgm $\mathbb{U}_x$ dans $\mathcal{C}_\xi$ que dans son 2δ-épaississement (de côté $2\delta + \xi < 3\delta$). Par construction, $\delta < \delta_0$ est assez petit pour pouvoir appliquer les inégalités des boîtes. On a donc

$$\mu_{\mathbb{U}_x}(C_\xi) \leq \mu_{\mathbb{U}_y}(C_\xi^\delta) \leq \mu_{\mathbb{U}_x}(C_\xi^{2\delta})$$

et il y a par conséquent autant de points de dgm $\mathbb{U}_y$ dans C_ξ^δ que de points de dgm $\mathbb{U}_x$ dans C_ξ. On peut associer injectivement à chaque point de dgm $\mathbb{U}_x$ un point de dgm $\mathbb{U}_y$, à distance plus petite que $\delta + \xi$.

Or, par la inégalité du lemme, on a $d_H(\text{dgm } \mathbb{U}_x, \text{dgm } \mathbb{U}_y) \leq \delta$, et donc pour tout point de dgm $\mathbb{U}_y$, soit il existe un carré C_δ centré en un certain point de dgm $\mathbb{U}_x$ qui le contient, soit il est à distance plus petite que δ de la diagonale. Dans le premier cas, il est en correspondance avec un point de dgm $\mathbb{U}_x$, et dans le deuxième cas il est laissé tout seul. De cette façon on obtient une correspondance partielle entre les points de dgm $\mathbb{U}_x$ et ceux de dgm $\mathbb{U}_y$ de coût au plus $\delta + \xi$. En faisant tendre ξ vers 0, on obtient finalement que $d_b(\text{dgm } \mathbb{U}_x, \text{dgm } \mathbb{U}_y) \leq \delta$.

□

On a maintenant tous les ingrédients dont on a besoin pour prouver l'inégalité de stabilité.

Théorème 5.12. (Inégalité de stabilité) *Soient $\mathbb{V}$, $\mathbb{W}$ deux modules de persistance q-modérés sur $\mathbb{R}$. Alors $d_b(\text{dgm } \mathbb{V}, \text{dgm } \mathbb{W}) \leq d_i(\mathbb{V}, \mathbb{W})$.*

Preuve. On ne va montrer ici que le cas où dgm $\mathbb{U}_x$ a un nombre fini de points pour tout $x \in [0, \varepsilon]$. Pour le cas général, on doit construire un prolongement de la mesure du rang sur les ensembles localement finis, cette construction peut être consultée dans la section 4.8 de [1].

On considère la famille d'intervalles $\left(]x - \delta_x, x + \delta_x[\cap[0,\varepsilon]\right)_{x\in[0,\varepsilon]}$ où δ_x est assez petit pour que $|x - y| \leq \delta_x$, ainsi par le lemme précédent $d_b(\text{dgm } \mathbb{U}_x, \text{dgm } \mathbb{U}_y) \leq d_H(\text{dgm } \mathbb{U}_x, \text{dgm } \mathbb{U}_y)$. Ils forment un recouvrement ouvert de $[0, \varepsilon]$ qui est compact, on peut donc en extraire un sous-recouvrement fini composé d'intervalles centrés en les points que l'on note $x_1 < ... < x_{k-1}$. On rajoute à ce recouvrement les intervalles centrés en $x_0 = 0$, en $x_k = \varepsilon$, et en $y_1, ..., y_k$ qui sont des points satisfaisant $x_i < y_{i+1} < x_{i+1}$. On prend deux éléments x, y consécutifs de cette suite (de la forme $x = x_i, y = y_i$ ou $x = x_{i+1}, y = y_i$), par construction $|x - y| < \delta_x$, donc par le lemme 5.11, $d_b(\text{dgm } \mathbb{U}_x, \text{dgm } \mathbb{U}_y) \leq |x - y|$. On conclut par l'inégalité triangulaire :

$$d_b(\text{dgm } \mathbb{V}, \text{dgm } \mathbb{W}) \leq \sum_{i=0}^{k} d_b(\text{dgm } \mathbb{U}_{x_i}, \text{dgm } \mathbb{U}_{y_i}) \leq \sum_{i=0}^{k} |x_i - y_i| = \varepsilon = d_i(\mathbb{V}, \mathbb{W}).$$

□

Ceci conclut la preuve de l'inégalité de stabilité.

5.2.2 Inégalité de stabilité inverse

La preuve de l'inégalité réciproque $d_b(\text{dgm } \mathbb{V}, \text{dgm } \mathbb{W}) \geq d_i(\mathbb{V}, \mathbb{W})$ est bien plus simple maintenant que l'on a montré celle de stabilité. On commence par faire la preuve dans le cas des modules de persistance décomposables en représentations d'intervalles.

Lemme 5.13. (Stabilité inverse pour les modules décomposables en intervalles) *Soient $\mathbb{V}$, $\mathbb{W}$ deux modules de persistance indexés sur $\mathbb{R}$ et décomposables en somme directe de représentations d'intervalles, alors $d_b(\text{dgm } \mathbb{V}, \text{dgm } \mathbb{W}) \geq d_i(\mathbb{V}, \mathbb{W})$.*

Preuve. On fixe $\varepsilon > d_b(\text{dgm } \mathbb{V}, \text{dgm } \mathbb{W})$, et on choisit une correspondance partielle M de (dgm $\mathbb{V}$, dgm $\mathbb{W}$) telle que $c(M) < \varepsilon$. Ces diagrammes induisent des décompositions de $\mathbb{V}$ et $\mathbb{W}$:

1. On commence en posant $J = \varnothing$.
2. Pour tout $(a, b) \in \text{dgm } \mathbb{V} \times \text{dgm } \mathbb{W}$ on rajoute un indice j à J et on définit $\mathbb{V}_j$ comme la représentation d'intervalle associée au point a, et $\mathbb{W}_j$ comme celle associée au point b.
3. Pour les points de dgm $\mathbb{V}$ qui n'ont pas de correspondant, on rajoute un indice j à J, et on définit $\mathbb{V}_j$ comme étant la représentation d'intervalle associée à ce point. On définit alors $\mathbb{W}_j$ comme étant la représentation nulle.
4. On fait de même pour les points de dgm $\mathbb{W}$ sans correspondant.

Ce qu'on a fait ici se réduit à ajouter des copies de la décomposition nulle (pour rendre la correspondance bijective) et à réindexer les décompositions en intervalle de $\mathbb{V}$ et $\mathbb{W}$ pour que deux intervalles avec le même indice soient en correspondance. De cette façon, on obtient des décompositions $\mathbb{V} = \oplus_{j\in J}\mathbb{V}_j$ et $\mathbb{W} = \oplus_{j\in J}\mathbb{W}_j$ telles que pour tout $j \in J$, les paires $(\mathbb{V}_j, \mathbb{W}_j)$ sont ε-entrelacées.

En effet, prenons une paire de représentations d'intervalles $\mathbb{V}_j = \mathbb{I}[p, q]$ et $\mathbb{W}_j = \mathbb{I}[p', q']$ tels que $|p - p'| < \varepsilon$ et $|q - q'| < \varepsilon$. On définit les morphismes $\phi_t : (V_j)_t \longrightarrow (W_j)_{t+\varepsilon}$, $\phi_t = id$ si $(v_i)_t = (w_i)_{t+\varepsilon}$ et nul sinon, et $\psi_t : (W_j)_t \longrightarrow (V_j)_{t+\varepsilon}$, $\psi_t = id$ si $(w_i)_t = (v_i)_{t+\varepsilon}$ et nul sinon. Ils font commuter les diagrammes de la définition 5.4, justement parce que $|p - p'| < \varepsilon$ et $|q - q'| < \varepsilon$. Et si $\mathbb{V}_j$ ou $\mathbb{W}_j$ est la représentation nulle, les morphismes identiquement nuls (qui d'ailleurs sont les seuls que l'on peut définir) conviennent. La somme directe sur $j \in J$ de ces entrelacements nous donne alors un ε-entrelacement entre $\mathbb{V}$ et $\mathbb{W}$, et on conclut en faisant tendre ε vers $d_b(\text{dgm } \mathbb{V}, \text{dgm } \mathbb{W})$. □

On montre enfin la stabilité inverse dans le cas général.

Théorème 5.14. (Stabilité inverse) *Soient* $\mathbb{V}$, $\mathbb{W}$ *deux modules de persistance* q*-modéré sur* $\mathbb{R}$, *alors* $d_b(\text{dgm } \mathbb{V}, \text{dgm } \mathbb{W}) \geq d_i(\mathbb{V}, \mathbb{W})$.

Preuve. Soit $\mathcal{P}_q$ l'ensemble des modules de persistance q-modérés sur $\mathbb{R}$. On pose

$$f : \left\{ \begin{array}{ccc} \mathcal{P}_q \times \mathcal{P}_q & \longrightarrow & [-\infty, \infty] \\ (\mathbb{V}, \mathbb{W}) & \mapsto & d_i(\mathbb{V}, \mathbb{W}) - d_b(\text{dgm } \mathbb{V}, \text{dgm } \mathbb{W}) \end{array} \right.$$

où l'on adopte la convention $\infty - \infty = 0$. On munit $\mathcal{P}_q \times \mathcal{P}_q$ de la distance en norme infinie (définie comme le maximum entre les distances d_i de chaque coordonnée). L'inégalité de stabilité prouvée dans la section précédente garantit la positivité de f. On a aussi par cette même inégalité que f est continue, en effet pour tout $\varepsilon > 0$, en posant $\delta = \frac{\varepsilon}{4} > 0$, si $d_i(\mathbb{V}, \mathbb{V}') \leq \delta$ et $d_i(\mathbb{W}, \mathbb{W}') \leq \delta$, alors

$$\begin{aligned} &|f(\mathbb{V}, \mathbb{W}) - f(\mathbb{V}', \mathbb{W}')| \\ &= |d_i(\mathbb{V}, \mathbb{W}) - d_b(\text{dgm } \mathbb{V}, \text{dgm } \mathbb{W}) - (d_i(\mathbb{V}', \mathbb{W}') - d_b(\text{dgm } \mathbb{V}', \text{dgm } \mathbb{W}')| \\ &\leq |d_i(\mathbb{V}, \mathbb{W})| + |d_i(\mathbb{V}', \mathbb{W}')| + |d_b(\text{dgm } \mathbb{V}, \text{dgm } \mathbb{W})| + |d_b(\text{dgm } \mathbb{V}, \text{dgm } \mathbb{W})| \\ &\leq 4\delta = \varepsilon. \end{aligned}$$

Or rappelons que par le théorème 2.8, un module de persistance de dimension finie en chaque point est décomposable en somme directe de représentations d'intervalles. Par conséquent, le lemme précédent nous garantit que f est identiquement nulle sur le sous-ensemble $\mathcal{P}_f \subseteq \mathcal{P}_q$ des modules de persistance sur $\mathbb{R}$ de dimension finie en chaque point. Il suffit donc pour conclure de montrer que $\mathcal{P}_f$ est dense dans $\mathcal{P}_q$. Soit donc $\mathbb{V}$ un module de persistance q-modéré, pour $i, j \in \mathbb{R}$, on note V_i l'espace vectoriel associé au point i et v_i^j le morphisme entre V_i et V_j. On fixe $\varepsilon > 0$ et on va construire un module $\mathbb{V}^\varepsilon$ de dimension finie en chaque point tel que $d_i(\mathbb{V}, \mathbb{V}^\varepsilon) < 2\varepsilon$. Pour tout $i \in \mathbb{R}$, on pose $V_i^\varepsilon = \text{im } v_{i-\varepsilon}^{i+\varepsilon}$ et on définit ainsi $\mathbb{V}^\varepsilon$ où les morphismes sont induits par ceux de $\mathbb{V}$. Comme $\mathbb{V}$ est q-modéré, $\mathbb{V}^\varepsilon$ est de dimension finie en chaque point. Et on a un ε-entrelacement entre $\mathbb{V}$ et $\mathbb{V}^\varepsilon$ en posant :

$$\phi_t : V_t \longrightarrow V_{t+\varepsilon}^\varepsilon = \text{im } v_t^{t+2\varepsilon} \quad \text{où} \quad \phi_t = v_t^{t+2\varepsilon}$$

et

$$\psi_t : V_t^\varepsilon = \text{im } v_{t-\varepsilon}^{t+\varepsilon} \longrightarrow V_{t+\varepsilon} \quad \text{l'inclusion canonique.}$$

Par conséquent, $d_i(\mathbb{V}, \mathbb{V}^\varepsilon) \leq \varepsilon < 2\varepsilon$ et l'argument détaillé plus haut conclut la preuve.

□

Deuxième partie

Estimation statistique

À partir d'une filtration q-modérée ou de dimension finie en chaque point d'un espace topologique, on a défini la notion de diagramme de persistance associé. On vient de voir que ces diagrammes sont stables et on se pose maintenant la question de l'estimation statistique. À partir d'un échantillonnage d'un espace topologique filtré, on veut construire un estimateur qui nous permet d'approximer le diagramme de persistance de la filtration. Dans cette partie, on va construire un tel estimateur et se poser la question de son optimalité.

6 Filtrations classiques

On commence par définir deux filtrations classiques qui nous seront utiles pour la suite.

Définition 6.1. Étant donné un recouvrement d'ouvert $X = \cup_{i\in I} U_i$, le *nerf du recouvrement* $\mathcal{U} = (U_i)_{i\in I}$, noté $\mathrm{Nerf}(\mathcal{U})$, est le complexe simplicial avec $\mathcal{U}$ comme ensemble de sommets et défini par

$$\sigma = [U_{i_0}, ..., U_{i_k}] \in \mathrm{Nerf}(\mathcal{U}) \iff \bigcap_{j=0}^{k} U_{i_j} \neq \emptyset.$$

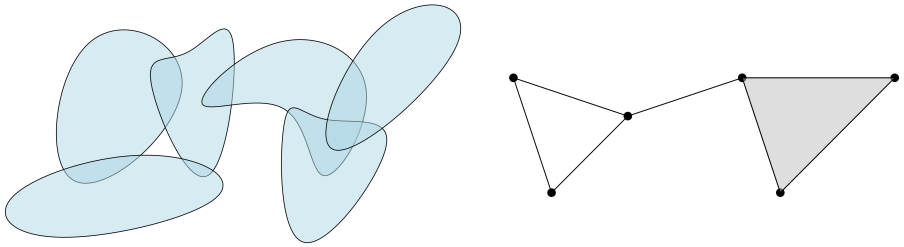


FIGURE 6.1 – Un recouvrement à 6 ouverts et son nerf associé.

Définition 6.2. Soient $\mathcal{P} \subset \mathbb{R}^d$ un ensemble fini de points, $\alpha > 0$. Le *complexe de Čech* $\check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, \alpha)$ est le nerf de l'union des boules centrées en les points de $\mathcal{P}$ et de rayon α.

$$\sigma = [p_0, ..., p_k] \in \check{\mathrm{C}}\mathrm{ech}(\mathcal{P}, \alpha) \iff \bigcap_{i=0}^{k} B(p_i, \alpha) \neq \emptyset.$$

Le *complexe de Vietoris-Rips* $\mathrm{Rips}(\mathcal{P}, \alpha)$ est quant à lui défini par

$$\sigma = [p_0, ..., p_k] \in \mathrm{Rips}(\mathcal{P}, \alpha) \iff ||p_i - p_j|| \leq \alpha \quad \forall i, j \in \{0, ..., k\}.$$

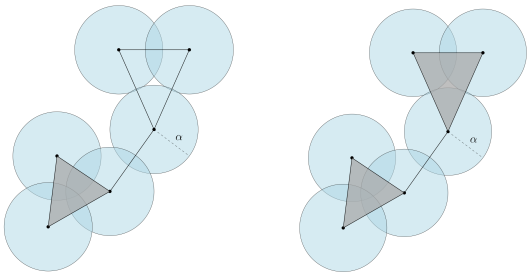


FIGURE 6.2 – Complexe de Čech (à gauche) et de Vietoris-Rips (à droite).

En faisant varier α de 0 à l'infini, la suite des complexes de Čech (resp. Vietoris-Rips) définit une filtration appelée *filtration du complexe de Čech* (resp. *filtration du complexe de Vietoris-Rips*). Pour simplifier la visualisation, les définitions ci-dessus sont données pour des ensembles finis de points. Cependant, on peut définir de la même manière les filtrations de Čech et de Vietoris-Rips pour des espaces métriques quelconques. Les complexes associés sont alors dit abstraits. Pour tout espace métrique compact $\mathbb{X}$, on notera dans la suite $\mathrm{Filt}(\mathbb{X})$ la filtration du complexe de Čech ou de Vietoris-Rips définie sur $\mathbb{X}$. La proposition ci-dessous nous assure que le diagramme de persistance de la filtration $\mathrm{Filt}(\mathbb{X})$ est bien défini. La preuve de cette proposition peut être trouvée dans la section 5.1 de [2], elle utilise en fait seulement la propriété de précompacité de $\mathbb{X}$.

Proposition 6.3. *Soit $\mathbb{X}$ un espace métrique compact, la filtration* $\mathrm{Filt}(\mathbb{X})$ *est q-modérée.*

7 Résultats probabilistes

On commence par définir le modèle statistique avec lequel on va travailler dans la suite.

Définition 7.1. Une distribution de probabilité μ est dite (a,b)-*standard à l'échelle* r_0 si pour tout $x \in \mathrm{supp}(\mu)$ et tout $0 \leq r \leq r_0$, $\mu\big(B(x,r)\big) \geq ar^b$. On notera dans la suite $\mathcal{P}_{M,a,b}$ l'ensemble des distributions de probabilité borélienne μ sur un espace métrique (M,ρ) tel que $\mathbb{X}_\mu = \mathrm{supp}(\mu)$ est compact et il existe $r_0 > 0$ tel que μ soit (a,b)-standard à l'échelle r_0.

On suppose dans toute la suite que l'on observe un échantillon i.i.d. $\mathbb{X}_n = \{X_1, ..., X_n\}$ tiré d'une distribution de probabilité inconnue $\mu \in \mathcal{P}_{M,a,b}$, définie sur un espace métrique (M,ρ). On rappelle que le support de μ, noté $\mathbb{X}_\mu$, est supposé compact. On rappelle également que $\mathrm{Filt}(\mathbb{X}_\mu)$ désigne la filtration du complexe de Čech ou de Vietoris-Rips définie sur $\mathbb{X}_\mu$. Notre objectif est de construire un estimateur optimal du diagramme de persistance de la filtration $\mathrm{Filt}(\mathbb{X}_\mu)$. Avant de construire cet estimateur et de montrer son optimalité, on commencera par quelques définitions et résultats techniques.

On rappelle tout d'abord la définition classique de la distance de Hausdorff.

Définition 7.2. Soient A, B deux sous-ensembles compacts de l'espace métrique (M,ρ). La *distance de Hausdorff* entre A et B est définie par

$$d_H(A,B) = \max\left\{\sup_{a\in A} d_B(a), \sup_{b\in B} d_A(b)\right\} = \max\left\{\sup_{a\in A}\inf_{b\in B}\rho(a,b), \sup_{b\in B}\inf_{a\in A}\rho(a,b)\right\}.$$

On va maintenant introduire une façon de mesurer la masse d'un ensemble compact basée sur ses recouvrements par des boules. Ceci fait sens puisque notre hypothèse sur le caractère (a,b)-standard des mesures utilisées va nous permettre d'établir des bornes sur les mesures des boules, et par suite sur celle du compact.

Définition 7.3. Soient $K \subseteq \mathbb{R}^d$ un compact et $r > 0$.

- Un *r-recouvrement* de K est un ensemble fini de points $R \subseteq \mathbb{R}^d$ tel que pour tout $x \in K$, il existe $y \in R$ tel que $x \in \overline{B(y,r)}$. L'ensemble des boules fermées de rayon r centrées en les points de R forme donc un recouvrement de K.
- Un *r-empaquetage* de K est un ensemble fini de points $P \subseteq \mathbb{R}^d$ tel que pour tous $x, y \in P$, $\overline{B(x,r)} \cap \overline{B(y,r)} = \varnothing$. Les boules fermées de rayon r centrées en les points de P sont donc toutes deux-à-deux disjointes et incluses dans K.
- Le *nombre de recouvrement* $\mathrm{cv}(K,r)$ est

$$\mathrm{cv}(K,r) = \min\big\{k \in \mathbb{N} \mid \text{il existe un } r\text{-recouvrement de } K \text{ de cardinal } k\big\}.$$

  Un recouvrement qui réalise ce minimum est appelé *recouvrement maximal*.

- Le *nombre d'empaquetage* $\mathrm{pk}(K,r)$ est le nombre maximal de boules fermées disjointes de $\mathbb{R}^d$ de rayon r qui peuvent être incluses dans K.

$$\mathrm{pk}(K,r) = \max\big\{k \in \mathbb{N} \mid \text{il existe un } r\text{-empaquetage de } K \text{ de cardinal } k\big\}.$$

  Tout empaquetage qui réalise ce maximum est appelé *empaquetage maximal*.

Notons que comme K est compact, il est de mesure bornée, et donc ces quantités sont bien définies. De plus, il existe toujours des recouvrements et des empaquetages maximaux (qui ne sont en général pas uniques). La proposition suivante relie les nombres d'empaquetage et de recouvrement.

Proposition 7.4. *Soient $K \subseteq \mathbb{R}^d$ un compact et $r > 0$. Alors*

$$\mathrm{pk}(K,2r) \leq \mathrm{cv}(K,2r) \leq \mathrm{pk}(K,r).$$

Preuve. On montre d'abord la première inégalité. Si, par l'absurde, R est un $2r$-recouvrement de K, et P un $2r$-empaquetage maximal tel que $|R| < |P|$, alors par le principe des tiroirs, il existe $y_1, y_2 \in P$ et $x \in R$ tels que $y_1, y_2 \in \overline{B(x,2r)}$. Mais alors $\overline{B(y_1,2r)} \cap \overline{B(y_2,2r)} \neq \varnothing$ ce qui contredit la définition d'un r-empaquetage.

Pour la deuxième inégalité, posons P un $2r$-empaquetage maximal. Alors P est aussi un r-recouvrement : en effet si, par l'absurde, il existait $x \in K$ tel que pour tout $y \in P$, $x \notin \overline{B(y,r)}$, on pourrait rajouter x à l'empaquetage.

□

On va maintenant donner une borne sur la masse du support d'une mesure de probabilité (a,b)-standard, qui dépend de son nombre de recouvrement.

Proposition 7.5. *Soit $\mu \in \mathcal{P}_{M,a,b}$ à l'échelle $r_0 > 0$. Alors, pour tout $r \leq 2r_0$,*

$$\mathrm{cv}\big(\mathrm{supp}(\mu), r\big) \leq \frac{2^b}{ar^b}\ .$$

Preuve. Soient $r \leq r_0$ et P un r-empaquetage maximal de $\mathrm{supp}(\mu)$. Comme la mesure est (a,b)-standard, pour tout $x \in P$, $\mu(B(x,r) \geq ar^b$. Par conséquent,

$$\begin{aligned}
\mathrm{pk}\big(\mathrm{supp}(\mu), r\big) \cdot ar^b = \sum_{x\in P} ar^b &\leq \sum_{x\in P} \mu\big(B(x,r)\big) \\
&\leq \sum_{x\in P} \mu\big(B(x,r)\big) \leq \mu\Big(\bigcup_{x\in P} B(x,r)\Big) \\
&\leq \mu(\mathbb{R}^d) = 1
\end{aligned}$$

Or, par la proposition précédente, $\mathrm{cv}(K,2r) \leq \mathrm{pk}(K,r)$. On en déduit donc l'énoncé. □

La proposition suivante est le résultat clé de cette section.

Proposition 7.6. *Soit $\mu \in \mathcal{P}_{M,a,b}$ à l'échelle r_0, alors pour tout $\varepsilon > 0$,*

$$\mathbb{P}\Big(d_H\big(\mathbb{X}_\mu, \mathbb{X}_n\big) > \varepsilon\Big) \leq \frac{4^b}{a\varepsilon^b} \exp\Big(-n\frac{a}{2^b}\varepsilon^b\Big).$$

Preuve. Soit $x \in \mathrm{supp}(\mu)$. Notre premier objectif est de borner la distance entre x et $\mathbb{X}_n$. Pour cela, on choisit $\delta > 0$, que l'on déterminera plus tard. Considérons un δ-recouvrement maximal R de $\mathrm{supp}(\mu)$. Il existe $y \in R$ tel que $x \in \overline{B(y,\delta)}$. Par conséquent,

$$\begin{aligned}
\min_{1\leq j\leq n} \|X_j - x\| &\leq \min_{1\leq j\leq n} \big(\|y - x\| + \|X_j - y\|\big) \\
&\leq \delta + \min_{1\leq j\leq n} \|X_j - y\| \\
&\leq \delta + \max_{z\in R} \min_{1\leq j\leq n} \|X_j - z\|.
\end{aligned}$$

Or, comme $\mathbb{X}_n \subseteq \mathrm{supp}(\mu)$,

$$d_H\big(\mathrm{supp}(\mu), \mathbb{X}_n\big) = \sup_{x' \in \mathrm{supp}(\mu)} \min_{1 \le j \le n} \|X_j - x'\|$$

et donc

$$d_H\big(\mathrm{supp}(\mu), \mathbb{X}_n\big) \le \delta + \max_{z \in R} \min_{1 \le j \le n} \|X_j - z\|.$$

Par conséquent,

$$\begin{aligned}\mathbb{P}\Big(d_H\big(\mathrm{supp}(\mu), \mathbb{X}_n\big) > r\Big) &\le \mathbb{P}\Big(\delta + \max_{z \in R} \min_{1 \le j \le n} \|X_j - z\| > r - \delta\Big) \\ &\le \sum_{z \in R} \mathbb{P}\Big(\min_{1 \le j \le n} \|X_j - z\| > r - \delta\Big).\end{aligned}$$

Or,

$$\mathbb{P}\Big(\min_{1 \le j \le n} \|X_j - y\| > r - \delta\Big) = \prod_{j=1}^{n} \mathbb{P}\Big(\|X_j - y\| > r - \delta\Big).$$

Donc, dès que $r - \delta > r_0$, on aura

$$\begin{aligned}\mathbb{P}\Big(\|X_j - y\| > r - \delta\Big) &= 1 - \mu\big(B(y, r - \delta)\big) \\ &\le 1 - a(r - \delta)^b \\ &\le \exp\big(-a(r - \delta)^b\big)\end{aligned}$$

où la dernière inégalité découle de la concavité du logarithme. On déduit de cette chaîne d'inégalités que

$$\mathbb{P}\Big(\min_{1 \le j \le n} \|X_j - y\| > r - \delta\Big) \le \prod_{i=1}^{n} \exp\big(-a(r - \delta)^b\big) = \exp\big(-na(r - \delta)^b\big).$$

Donc,

$$\mathbb{P}\Big(d_H\big(\mathrm{supp}(\mu), \mathbb{X}_n\big) > r\Big) \le \sum_{z \in R} \exp\big(-na(r - \delta)^b\big) = \mathrm{cv}\big(\mathrm{supp}(\mu), \delta\big) \cdot \exp(-na(r - \delta)^b).$$

Et par la proposition précédente, si $\delta \le 2r_0$, on aura

$$\mathbb{P}\Big(d_H\big(\mathrm{supp}(\mu), \mathbb{X}_n\big) > r\Big) \le \frac{2^b}{a\delta^b} \exp\big(-na(r - \delta)^b\big).$$

L'énoncé est montré en posant $\delta = r/2$.

□

On en déduit en particulier le corollaire suivant :

Corollaire 7.7. *Pour tout $\alpha > 0$, il existe une constante $C > 0$ (dépendante de a, b et α) telle que pour tout $n \in \mathbb{N}$ tel que $\big(C\frac{\log n}{n}\big)^{1/b} \le 2r_0$,*

$$\mathbb{P}\left(d_H\big(\mathrm{supp}(\mu), \mathbb{X}_n\big) \le C\Big(\frac{\log n}{n}\Big)^{1/b}\right) \ge 1 - n^{-\alpha}.$$

Pour une mesure (a, b)-standard, la vitesse de convergence d'un nuage de points vers son support est donc de l'ordre de $\Big(\log(n)/n\Big)^{1/b}$.

8 Estimation du diagramme de persistance

Nous avons défini précédemment la distance de Hausdorff pour deux sous-espaces compacts d'un espace métrique. Il se peut néanmoins que l'on veuille comparer deux espaces métriques compacts quelconques. On introduit donc ci-dessous la distance de Gromov-Hausdorff.

Définition 8.1. Soient (M_1, ρ_1), (M_2, ρ_2) deux espaces métriques compacts. La *distance de Gromov-Hausdorff* entre M_1 et M_2, noté $d_{GH}(M_1, M_2)$, est l'infimum des réels $r \geq 0$ tel qu'il existe un espace métrique (M, ρ) et deux compacts $K_1, K_2 \subset M$ isométriques à M_1 et M_2 avec $d_H(K_1, K_2) \leq r$.

On déduit de cette définition que pour tous sous-espaces compacts A, B d'un espace métrique (M, ρ), on a $d_{GH}(A, B) \leq d_H(A, B)$. La preuve de la proposition suivante peut être trouvée dans la section 4.2 de [2].

Proposition 8.2. *Soient $\mathbb{X}, \mathbb{Y}$ des espaces métriques, $\varepsilon > d_{GH}(\mathbb{X}, \mathbb{Y})$ et $p \in \mathbb{N}$, alors les modules de persistance $\mathcal{H}_p\big(\mathrm{Filt}(\mathbb{X})\big)$ et $\mathcal{H}_p\big(\mathrm{Filt}(\mathbb{Y})\big)$ sont ε-entrelacé. En particulier,*

$$d_i\Big(\mathcal{H}_p\big(\mathrm{Filt}(\mathbb{X})\big), \mathcal{H}_p\big(\mathrm{Filt}(\mathbb{Y})\big)\Big) \leq 2d_{GH}(\mathbb{X}, \mathbb{Y}).$$

On est maintenant prêt à construire notre estimateur du diagramme de persistance. On se rappelle notre objectif : on dispose d'un échantillon i.i.d. $\mathbb{X}_n = \{X_1, ..., X_n\}$ tiré d'une distribution de probabilité inconnue $\mu \in \mathcal{P}_{M,a,b}$ de support $\mathbb{X}_\mu$ compact, définie sur un espace métrique (M, ρ). Notre objectif est de construire un estimateur optimal du diagramme de persistance de la filtration $\mathrm{Filt}(\mathbb{X}_\mu)$. On va considérer pour cela l'estimateur $\mathrm{dgm}\big(\mathrm{Filt}(\mathbb{X}_n)\big)$. Le choix de cet estimateur se justifie d'une part par la proposition 7.6 et d'autre part par le théorème suivant, lequel est une conséquence directe du théorème d'isométrie, de la proposition 6.3 et de la proposition précédente.

Théorème 8.3. *Soient $\mathbb{X}, \mathbb{Y}$ deux espaces métriques compacts, alors*

$$d_b\Big(\mathrm{dgm}\big(\mathrm{Filt}(\mathbb{X})\big), \mathrm{dgm}\big(\mathrm{Filt}(\mathbb{Y})\big)\Big) \leq 2d_{GH}\big(\mathbb{X}, \mathbb{Y}\big).$$

On a maintenant tous les outils pour démontrer les deux théorèmes suivants qui établissent une borne supérieure sur la vitesse de convergence de notre estimateur du diagramme de persistance.

Théorème 8.4. *Soit $\mu \in \mathcal{P}_{M,a,b}$, alors pour tout $\varepsilon > 0$,*

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

Preuve. On utilise le théorème précédent ainsi que la proposition 7.6

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

□

Théorème 8.5. *Soit $\mu \in \mathcal{P}_{M,a,b}$, alors pour n suffisament grand, il existe $\lambda_{a,b} \in \mathbb{R}$ tel que*

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

Preuve. En utilisant le théorème de Fubini, on peut écrire l'espérance sous la forme

$$\int_0^\infty \mathbb{P}\Big(d_b\Big(\mathrm{dgm}\big(\mathrm{Filt}(\mathbb{X}_\mu)\big), \mathrm{dgm}\big(\mathrm{Filt}(\mathbb{X}_n)\big)\Big) > \varepsilon\Big) d\varepsilon.$$

On peut alors appliquer le théorème précédent pour conclure. Le détail des calculs peut être trouvé dans la section 6.1 de [8].

□

Définition 8.6. Soient Q, Q' deux distributions de probabilité sur un ensemble mesurable $(\mathcal{X}, \mathcal{A})$, la *distance en variation totale* entre Q et Q' est définie par

$$d_{VT}(Q, Q') = \sup_{A \in \mathcal{A}} |Q(A) - Q'(A)|.$$

Cela définit une distance sur l'espace des mesures de probabilité sur $(\mathcal{X}, \mathcal{A})$. Pour q, q' les densités associées, on introduit aussi la notation $(q \wedge q')(x) = \min\{q(x), q'(x)\}$.

On va maintenant montrer que notre estimateur est optimal à un terme logarithmique près. Pour cela, on va avoir besoin du lemme de Le Cam que l'on démontre ci-dessous.

Proposition 8.7. *Soient Q, Q' deux distributions de probabilité sur un ensemble mesurable $(\mathcal{X}, \mathcal{A})$, alors pour toute mesure ν dominant Q et Q' et q, q' leur densité associée $\big($c'est-à-dire $dQ(x) = q(x)d\nu(x)$ et $dQ'(x) = q'(x)d\nu(x)\big)$, on a*

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

Preuve. En posant $A_0 = \{q \geq q'\}$, on a

$$0 = \int_{\mathcal{X}} (q - q') d\nu = \int_{A_0} (q - q') d\nu - \int_{A_0^c} (q - q') d\nu,$$

$$\int_{\mathcal{X}} |q - q'| d\nu = \int_{A_0} (q - q') d\nu + \int_{A_0^c} (q' - q) d\nu = 2\int_{A_0} (q - q') d\nu.$$

On en déduit donc que

$$d_{VT}(Q, Q') \geq Q(A_0) - Q'(A_0) = \frac{1}{2}\int_{\mathcal{X}} |q - q'| d\nu.$$

D'autre part, pour tout $A \in \mathcal{A}$,

$$\begin{aligned} |Q(A) - Q'(A)| &= \left| \int_{A \cap A_0} (q - q') d\nu + \int_{A \cap A_0^c} (q - q') d\nu \right| \\ &\leq \max\left\{ \int_{A \cap A_0} (q - q') d\nu, \int_{A \cap A_0^c} (q' - q) d\nu \right\} \\ &\leq \frac{1}{2}\int_{\mathcal{X}} |q - q'| d\nu. \end{aligned}$$

La deuxième égalité découle du fait que $|q - q'| = q + q' - 2(q \wedge q')$. □

Lemme 8.8. (Le Cam) *Soient $\mathcal{Q}$ un ensemble de distribution de probabilité, (Θ, ℓ) un espace métrique et $\theta : \mathcal{Q} \to \Theta$ un paramètre d'interêt. Alors on a pour tous $Q, Q' \in \mathcal{Q}$*

$$\inf_{\hat{\theta}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n}\left[\ell\big(\theta(Q), \hat{\theta}_n\big)\right] \geq \frac{1}{2}\ell\big(\theta(Q), \theta(Q')\big)\big(1 - d_{TV}(Q, Q')\big)^n$$

où $\hat{\theta}_n = \hat{\theta}_n(X_1, ...X_n)$ parcourt l'ensemble des fonction mesurable $\hat{\theta}_n : \mathcal{X}^n \to \Theta$ basée sur un échantillon $X_1, ..., X_n$ i.i.d.

Preuve. Soient ν une mesure dominant Q et Q' et q, q' les densités associées. Alors pour tout $\hat{\theta}_n : \mathcal{X}^n \to \Theta$ mesurable, on a

$$\begin{aligned}
\sup_{Q\in\mathcal{Q}} \mathbb{E}_{Q^n}\ell\big(\theta(Q),\hat{\theta}_n\big) &\geq \frac{1}{2}\Big(\mathbb{E}_{Q^n}\ell\big(\theta(Q),\hat{\theta}_n\big) + \mathbb{E}_{Q'^n}\ell\big(\theta(Q'),\hat{\theta}_n\big)\Big) \\
&= \frac{1}{2}\int_{\mathcal{X}^n}\Big(\ell\big(\theta(Q),\hat{\theta}_n\big)q^{\otimes n} + \ell\big(\theta(Q'),\hat{\theta}_n\big)q'^{\otimes n}\Big)\,d\nu^{\otimes n} \\
&\geq \frac{1}{2}\int_{\mathcal{X}^n}\Big(\ell\big(\theta(Q),\hat{\theta}_n\big) + \ell\big(\theta(Q'),\hat{\theta}_n\big)\Big)\,q^{\otimes n}\wedge q'^{\otimes n}d\nu^{\otimes n} \\
&\geq \frac{1}{2}\ell\big(\theta(Q),\theta(Q')\big)\prod_{i=1}^{n}\int_{\mathcal{X}} q(x_i)\wedge q'(x_i)d\nu(x_i) \\
&= \frac{1}{2}\ell\big(\theta(Q),\theta(Q')\big)\big(1-d_{VT}(Q,Q')\big)^n.
\end{aligned}$$

□

Théorème 8.9. *Supposons qu'il existe un point non-isolé $x \in M$, soit $(x_n)_{n\in\mathbb{N}} \subseteq M\setminus\{x\}$ tel que $\rho(x,x_n) \leq (an)^{-1/b}$. Alors pour tout estimateur $\widehat{\mathrm{dgm}}_n = \widehat{\mathrm{dgm}}_n(X_1,...,X_n)$, on a*

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

Preuve. On va appliquer le lemme de Le Cam au modèle $\mathcal{Q} = \mathcal{P}_{M,a,b}$ et au paramètre d'interêt $\theta : \mu \mapsto \mathrm{dgm}\big(\mathrm{Filt}(\mathbb{X}_\mu)\big)$ dans l'espace (Θ,ℓ) des diagrammes de persistance de module q-modéré avec $\ell = d_b$. On considère les distributions de Dirac $\mu_0 = \delta_x$ et $\mu_{1,n} = \frac{1}{n}\delta_{x_n} + (1-1\frac{1}{n})\mu_0$, de support respectif $\mathbb{X}_0 = \{x\}$ et $\mathbb{X}_{1,n} = \{x, x_n\}$. Pour pouvoir appliquer le lemme de Le Cam, il faut démontrer que ces deux distributions sont dans le modèle $\mathcal{P}_{M,a,b}$. On a clairement $\mu_0 \in \mathcal{P}_{M,a,b}$, et pour $\mu_{1,n}$, on remarque que pour tous $n \geq 2$ et $r \leq \rho(x,x_n)$, on a

$$\mu_{1,n}\big(B(x,r)\big) = 1 - \frac{1}{n} \geq \frac{1}{2} \geq \frac{1}{2\rho(x,x_n)^b}r^b \geq ar^b,$$

$$\mu_{1,n}\big(B(x_n,r)\big) = \frac{1}{n} = \frac{1}{n\rho(x,x_n)^b}r^b \geq ar^b,$$

et pour $r > \rho(x,x_n)$, on a $\mu_{1,n}\big(B(x,r)\big) = \mu_{1,n}\big(B(x_n,r)\big) = 1$. Donc pour tous $r > 0$ et $x \in \mathbb{X}_{1,n}$, on a

$$\mu_{1,n}\big(B(x,r)\big) \geq \min\{ar^b, 1\},$$

et donc cela prouve que $\mu_{1,n} \in \mathcal{P}_{M,a,b}$.

La mesure μ_0 est dominée par $\mu_{1,n}$ et a comme densité associée $p_{0,n} = \frac{n}{n-1}\mathbb{1}_{\{x\}}$.

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

on a donc $\big(1-d_{VT}(\mu_0,\mu_{1,n})\big)^n = (1-\frac{1}{n})^n \to e^{-1}$.

On doit maintenant calculer $d_b\Big(\mathrm{dgm}\big(\mathrm{Filt}(\mathbb{X}_0)\big),\mathrm{dgm}\big(\mathrm{Filt}(\mathbb{X}_{1,n})\big)\Big)$. Ces diagrammes ne sont non triviaux que pour l'homologie de dimension 0. On a d'une part $\mathrm{dgm}_0\big(\mathrm{Filt}(\mathbb{X}_0)\big) = \{(0,+\infty)\}$ et d'autre part $\mathrm{dgm}_0\big(\mathrm{Filt}(\mathbb{X}_{1,n})\big) = \{(0,+\infty),(0,\rho(x,x_n))\}$. Donc

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

Le lemme de Le Cam permet de conclure.

□

L'estimateur $\mathrm{dgm}\big(\mathrm{Filt}(\mathbb{X}_n)\big)$ est donc optimal sur l'espace $\mathcal{P}_{M,a,b}$ à un terme logarithmique près. Cela n'est en fait vrai que si l'on peut trouver un point non isolé $x \in M$ et une suite $(x_n)_{n\in\mathbb{N}} \subseteq M \setminus \{x\}$ tel que $\rho(x, x_n) \leq (an)^{-1/b}$, mais cette condition est vérifiée dans $\mathbb{R}^d$.

Pour conclure ce mémoire, on commentera brièvement les aspects statistiques du diagramme de persistance. Cet objet, que nous avons défini de manière algébrique via les modules de persistance, possède une propriété de stabilité qui garantit sa robustesse vis-à-vis de perturbations métriques. C'est cette stabilité qui nous a conduits à étudier la question de son estimation statistique. À partir d'un échantillon de données issues d'une loi de probabilité (a, b)-standard à support compact, nous avons construit un estimateur du diagramme de persistance d'une filtration associée au support. Cet estimateur s'est avéré optimal, à un terme logarithmique près.

Néanmoins, cette approche se heurte à une difficulté bien connue en géométrie et en statistique : *le fléau de la dimension*. Ce phénomène, introduit par Richard Bellman au début des années 1960, désigne l'inefficacité croissante des méthodes d'estimation à mesure que la dimension augmente. Dans notre cadre, la vitesse de convergence de l'estimateur dépend fortement de la dimension b de la loi $\mu \in \mathcal{P}_{M,a,b}$. Plus cette dimension est élevée, plus la convergence est lente. Bien que le diagramme de persistance soit un objet vivant dans un espace de dimension 2, muni d'une métrique fixe, la complexité statistique de son estimation dépend fortement de la dimension de la loi échantillonnée.

Références

[1] Frédéric Chazal, Vin de Silva, Marc Glisse et Steve Oudot. *The structure and stability of persistence modules*. 2013. URL : https://arxiv.org/pdf/1207.3674.

[2] Frédéric Chazal, Vin de Silva et Steve Oudot. *Persistence stability for geometric complexes*. 2013. URL : https://arxiv.org/pdf/1207.3885.

[3] William Crawley-Boevey. *Decomposition of pointwise finite-dimensional persistence modules*. 2014. URL : https://arxiv.org/pdf/1210.0819.

[4] Steve Oudot. *Persistence theory : From quiver representations to data analysis*. Providence, RI : American Mathematical Society., 2015.

[5] Justin Curry. *The Fiber of the Persistence Map for Functions on the Interval*. 2019. URL : https://arxiv.org/pdf/1706.06059.

[6] Henry Hale. *An exposition of a proof of Gabriel's theorem*. 2021. URL : https://math.uchicago.edu/~may/REU2021/REUPapers/Hale.pdf.

[7] Julien Tierny. *original 2D data, 3D terrain representation and persistence diagram*. [illustration page 1]. 2022. URL : https://julien-tierny.github.io/stuff/openPositions/internship2022.pdf.

[8] Eddie Aamari. *Persistent homology*. URL : https://www.math.ens.psl.eu/~eaamari/teaching/2023-2024/M2_Jussieu/Lesson%205.pdf.