

RAPPORT DE STAGE

Analyse topologique des données

Raphaël TINARRAGE
M2 MSV, Orsay

DATASHAPE
INRIA Paris-Saclay

30 janvier 2021

Table des matières

Introduction	3
1 Algèbre persistante	4
1.1 Exemple introductif	4
1.2 Modules de persistance	6
1.3 Entrelacements	8
1.4 Décomposition des modules de persistance	10
1.4.1 Définitions	10
1.4.2 Décomposition des modules de type fini sur \mathbb{N}	11
1.4.3 Décomposition des modules modérés sur \mathbb{R}^+	14
1.4.4 Décomposition des modules finis partout sur \mathbb{R}	15
1.5 Diagrammes de persistance	16
1.5.1 Définition	16
1.5.2 Stabilité	17
1.6 Le théorème d'isométrie	18
1.6.1 Cas des modules modérés	18
1.6.2 Généralisation aux modules q-modérés	18
2 Espaces topologiques filtrés	20
2.1 Filtrations	20
2.2 Fonctions de Morse	21
2.3 Fonctions distance	22
2.3.1 Axe médian	22
2.3.2 Cellules de Voronoï	23
2.3.3 Gradient généralisé	23
2.4 Fonctions presque-distance	25
2.5 Stabilité des modules	26
2.5.1 Régularité	26
2.5.2 Stabilité	26
3 Complexes simpliciaux filtrés	28
3.1 Filtrations	28
3.2 Algorithme de persistance	29
3.2.1 Calcul des nombres de Betti	29
3.2.2 Nombres de Betti persistants	30
3.3 Complexe de Čech	33
3.3.1 Théorème du nerf persistant	33
3.3.2 Complexe de Vietoris-Rips	34
3.4 Stabilité	36
3.4.1 Applications ϵ -simpliciales	36
3.4.2 Correspondances	37
3.4.3 Stabilité en Gromov-Hausdorff	37
4 Distance à la mesure	38
4.1 Introduction	38
4.2 Définition	39
4.3 Régularité et stabilité	40
4.3.1 Lien avec les distances de Wasserstein	40
4.3.2 La DTM est une presque-distance	40

4.3.3	Stabilité	41
4.4	Calcul et approximation	42
4.4.1	Décomposition barycentrique	42
4.4.2	Complexe de Čech pondéré	43
4.5	Simulation	45
5	D'autres structures pour l'inférence topologique	46
5.1	Le problème de la stabilité	46
5.2	Structure \mathbb{W}	47
5.2.1	Définition	47
5.2.2	Contraction d'arêtes	47
5.2.3	Approximation et stabilité	48
5.3	Structure \mathbb{F}	50
5.3.1	Définition	50
5.3.2	Approximation et stabilité	50
5.3.3	Complexe de clique de \mathbb{F}	51
5.4	Simulations	52
A	Formulaire d'homologie dans $\mathbb{Z}/2\mathbb{Z}$	53
A.1	Groupes d'homologie	53
A.2	Homologie singulière	53
A.3	Homologie simpliciale	54
	Références	56

Introduction

Dans le domaine de l'analyse de données, nous avons souvent affaire à des ensembles de données discrètes, que l'on peut interpréter comme des nuages de points dans des espaces euclidiens de grande dimension. Pour les traiter, l'analyse topologique des données repose sur le principe qu'il existe des informations de nature topologique et géométrique à extraire de ces nuages. De par les contraintes mises en jeu, nous pouvons présumer que ces données se concentrent sur des sous-espaces de petite dimension. Ainsi, si le nuage est suffisamment dense, il est envisageable de les estimer.

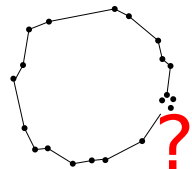
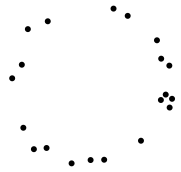
Parmi les outils mathématiques à disposition pour l'étude des espaces topologiques, l'homologie se révèle ici être d'une pertinence particulière : elle permet de caractériser ces espaces, et prend des formes implémentables du point de vue algorithmique.

L'objet de ce rapport de stage est de présenter la théorie de la persistance, ou comment extraire d'un échantillon de points l'homologie de l'espace sous-jacent. Nous introduirons ensuite la distance à la mesure, qui permet de traiter un nuage de points comme une mesure de probabilité.

1 Algèbre persistante

1.1 Exemple introductif

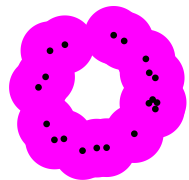
Soit X un nuage fini de points, par exemple tirés sur le cercle $\mathbb{S}^1 \subset \mathbb{R}^2$. Disons même que les observations soient bruitées, si bien que $X \not\subset \mathbb{S}^1$. En analyse topologique des données, on cherche l'homologie de l'espace sous-jacent (\mathbb{S}^1) à partir des observations (X). Dans ce mémoire, nous ne nous intéresserons qu'à l'homologie dans le corps $k = \mathbb{Z}/2\mathbb{Z}$. Nous renvoyons à l'annexe A pour les définitions et conventions utilisées.



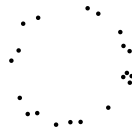
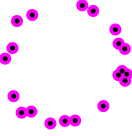
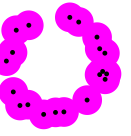
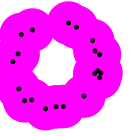
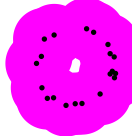
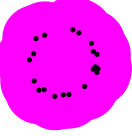
Une première idée consiste à construire sur X un complexe simplicial qui serait homéomorphe à \mathbb{S}^1 . Toutefois, la reconstruction de variété est une tâche difficile. La présence de bruit complique encore plus le problème. En fait, puisque l'on s'intéresse seulement à son homologie, nous n'avons pas besoin de retrouver l'espace sous-jacent à homéomorphisme près, mais seulement à type d'homotopie près. Par exemple, un petit grossissement du cercle est

dans sa classe d'homotopie.

Posons alors sur chacun des points de X une boule de rayon t . L'espace obtenu est $X^t := \bigcup_{x \in X} B(x, t)$, le voisinage tubulaire de X . Si le rayon est bien choisi, on voit sur la figure que \mathbb{S}^1 est un rétract par déformation de X^t , et ces deux espaces sont bien du même type d'homotopie. Ainsi, si nous sommes capables de calculer l'homologie de X^t , nous obtenons aussi celle de \mathbb{S}^1 , et le problème est résolu.



En pratique, le problème se pose lorsque nous ne savons pas choisir a priori le paramètre t . Dans notre exemple, s'il est trop petit, le voisinage tubulaire est composé de plusieurs parties connexes; s'il est trop grand, il est homéomorphe à un disque. L'idée de la persistance est de calculer l'homologie de ces voisinages tubulaires X^t pour toutes les valeurs de $t \in \mathbb{R}^+$, et de suivre au fur et à mesure leur évolution.

t	0	0,1	0,3	0,5	...	0,8	1
X^t					...		
$H_0(X^t)$	k^{20}	k^{14}	k^2	k	...	k	k
$H_1(X^t)$	0	0	0	k	...	k	0

Nous avons ici affaire à une suite croissante de sous-espaces topologiques $\{X^t\}_{t \in \mathbb{R}^+}$, avec des inclusions :

$$\dots \subset X^r \subset X^s \subset X^t \subset \dots$$

L'homologie (singulière) de ces espaces est une suite de groupes $H_n(X^t)$, $n \in \mathbb{N}$. Les inclusions précédentes induisent des morphismes entre ces groupes :

$$\dots \rightarrow H_*(X^r) \rightarrow H_*(X^s) \rightarrow H_*(X^t) \rightarrow \dots$$

(où l'étoile signifie que l'on observe un groupe d'homologie quelconque).

Mieux encore, par fonctorialité de l'homologie, la commutativité du diagramme au niveau des inclusions se traduit en une commutativité au niveau homologique :

$$\begin{array}{ccccc} X^r & \hookrightarrow & X^s & \hookrightarrow & X^t \\ & \searrow & & \nearrow & \\ & & & & \end{array} \quad \text{donne} \quad \begin{array}{ccccc} H_*(X^r) & \longrightarrow & H_*(X^s) & \longrightarrow & H_*(X^t) \\ & \searrow & & \nearrow & \\ & & & & \end{array}$$

La structure algébrique que nous venons de construire s'appelle un module de persistance. Dans cette structure, nous sommes capables de suivre l'évolution des cycles de X^t , leur naissance et leur mort. En effet, les applications $H_*(X^s) \rightarrow H_*(X^t)$ indiquent si un cycle de $H_*(X^s)$ est encore présent dans $H_*(X^t)$. Il se trouve que tout ce problème se traite de manière purement algébrique, et c'est pourquoi nous ne reparlerons presque plus de topologie dans ce chapitre.

1.2 Modules de persistance

Les définitions et résultats présentés dans les parties suivantes proviennent de [1].

Dans ce mémoire, nous n'aurons besoin de définir les modules de persistance que sur $T = \mathbb{N}$, \mathbb{R}^+ ou \mathbb{R} .

Définition 1.2.1. Un module de persistance (indexé par T) est la donnée d'une famille de k -espaces vectoriels $\mathbb{V} = (V_t)_{t \in T}$ et d'applications linéaires $(v_s^t : V_s \rightarrow V_t)_{s \leq t \in T}$ tels que

- $\forall t \in T, v_t^t$ est la fonction identité sur V_t , et
- $\forall s \leq t \leq u \in T, v_t^u \circ v_s^t = v_s^u$.

La dernière condition se traduit par le diagramme commutatif suivant, pour tout $\forall s \leq t \leq u \in T$:

$$\begin{array}{ccc} V_r & \xrightarrow{v_r^t} & V_t \\ & \searrow v_r^s \quad \nearrow v_s^t & \\ & V_s & \end{array}$$

Si $T = \mathbb{N}$ ou \mathbb{Z} , les applications v_s^t sont entièrement déterminées par les $v_n^{n+1}, n \in T$.

Remarque 1.2.2. Dans le langage des catégories, un module de persistance est un foncteur de l'ensemble totalement ordonné T vers la catégorie des espaces vectoriels.

Exemple 1.2.3. Soit $(K^t)_{t \in \mathbb{R}}$ une famille croissante de sous-ensembles d'un même espace topologique. Comme nous l'avons vu dans la partie 1.1, les inclusions induisent des diagrammes commutatifs entre les groupes d'homologie, qui font de $H_*(K^t)_{t \in \mathbb{R}}$ un module de persistance.

Définition 1.2.4. Soit $\epsilon \in T^+$. Si \mathbb{U} et \mathbb{V} sont deux modules de persistance, on appelle morphisme de degré ϵ entre \mathbb{U} et \mathbb{V} une famille de morphismes $\Phi = (\phi_t : U_t \rightarrow V_{t+\epsilon})_{t \in T}$ telle que le diagramme suivant commute en tout $s \leq t \in T$:

$$\begin{array}{ccc} U_s & \xrightarrow{u_s^t} & U_t \\ \downarrow \phi_s & & \downarrow \phi_t \\ V_{s+\epsilon} & \xrightarrow{v_{s+\epsilon}^{t+\epsilon}} & V_{t+\epsilon} \end{array}$$

Si $\mathbb{U} = \mathbb{V}$, Φ est appelé un endomorphisme. L'ensemble des endomorphismes de degré 0 de \mathbb{U} est une k -algèbre (pour la somme et la composition des applications linéaires).

Exemple 1.2.5. Soient $(K^t)_{t \in \mathbb{R}}$ et $(L^t)_{t \in \mathbb{R}}$ deux familles croissantes de sous-ensembles d'un même espace topologique. On suppose qu'il existe $\epsilon \geq 0$ tel que, pour tout $t \in \mathbb{R}$, on ait des inclusions $K^t \subset L^{t+\epsilon} \subset K^{t+2\epsilon}$. Les inclusions induisent un diagramme commutatif entre les groupes d'homologie :

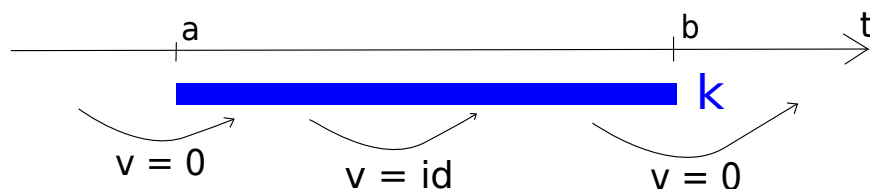
$$\begin{array}{ccc} K^s & \hookrightarrow & K^t \\ \downarrow & & \downarrow \\ L^{s+\epsilon} & \hookrightarrow & L^{t+\epsilon} \end{array} \quad \text{donne} \quad \begin{array}{ccc} H_*(K^s) & \longrightarrow & H_*(K^t) \\ \downarrow & & \downarrow \\ H_*(L^{s+\epsilon}) & \longrightarrow & H_*(L^{t+\epsilon}) \end{array}$$

Autrement dit, si l'on appelle \mathbb{U} et \mathbb{V} les modules de persistance induits par l'homologie singulière des familles croissantes (K_t) et (L_t) , alors les inclusions $(\phi_t : K^t \hookrightarrow K^{t+\epsilon})_{t \in \mathbb{R}}$ induisent un ϵ -morphisme entre \mathbb{U} et \mathbb{V} .

Modules-intervalle. Comme nous le verrons à la section 1.4, les modules-intervalle sont les briques élémentaires des modules de persistance. Pour les construire, on se donne un intervalle $I \subset T$, et on définit le module de persistance $\mathbb{I}(I)$ par :

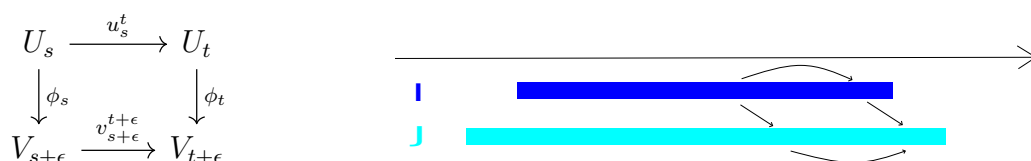
- Les espaces vectoriels, pour tout $t \in T$, $\begin{cases} V_t = k & \text{si } t \in I \\ = 0 & \text{sinon.} \end{cases}$
- Les morphismes, pour $s \leq t \in T$, $\begin{cases} v_s^t = \text{id} & \text{si } s, t \in I \\ = 0 & \text{sinon.} \end{cases}$

On peut dessiner ce module sous forme de code-barre. La figure suivante représente le module intervalle $\mathbb{I}([a, b])$ sur \mathbb{R} .



Entre deux modules intervalles $\mathbb{I}(I)$ et $\mathbb{I}(J)$ sur \mathbb{R} , on peut toujours définir, pour tout $\epsilon \geq 0$, un morphisme $\Phi : \mathbb{U} \rightarrow \mathbb{V}$ de degré ϵ par $\begin{cases} \phi_t \text{ est l'identité si } t \in I \text{ et } t + \epsilon \in J \\ \phi_t = 0 \text{ sinon.} \end{cases}$

La propriété de commutativité de Φ s'illustre par la figure suivante :



Proposition 1.2.6. Si $I \neq \emptyset$, le groupe des endomorphismes d'un module intervalle $\mathbb{I}(I)$ est isomorphe à k (en tant que k -algèbre).

Démonstration. Soit Φ un endomorphisme de $\mathbb{I}(I)$. En chaque $t \in I$, $\phi_t : V_t \rightarrow V_t$ est un morphisme du k -espace vectoriel $V_t = k$, et s'écrit donc $x \mapsto \lambda x$, avec $\lambda \in k$. Or la propriété de commutativité d'un morphisme de modules de persistance nous donne que la constante λ ne dépend pas de $t \in T$. On peut ainsi identifier Φ et λ . \square

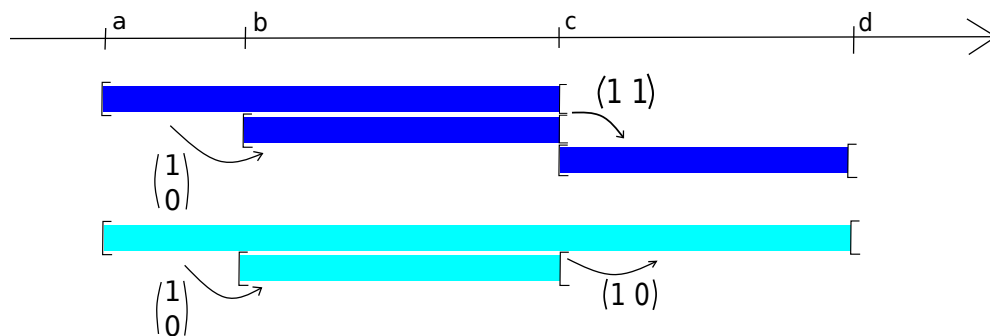
1.3 Entrelacements

Nous considérons dans cette partie des modules sur $T = \mathbb{R}$.

Définition 1.3.1. On dit que deux modules de persistance \mathbb{U} et \mathbb{V} sont isomorphes s'il existe un isomorphisme $\Phi : \mathbb{U} \rightarrow \mathbb{V}$ de degré 0, c'est à dire une famille d'isomorphismes $(\phi_t : U_t \rightarrow V_t)_{t \in T}$ telle qu'on ait des diagrammes commutatifs :

$$\begin{array}{ccc} U_s & \xrightarrow{u_s^t} & U_t \\ \downarrow \phi_s & & \downarrow \phi_t \\ V_s & \xrightarrow{v_s^t} & V_t \end{array}$$

Exemple 1.3.2. Considérons les deux modules de persistance suivants :



où les matrices représentent des applications linéaires.

Soit la famille d'isomorphismes Φ telle que :

- Pour $a \leq t < b$, $\phi_1 = \text{id} : k \rightarrow k$,
- pour $b \leq t < c$, $\phi_2 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} : k^2 \rightarrow k^2$,
- pour $c \leq t \leq d$, $\phi_3 = \text{id} : k \rightarrow k$,
- et $\phi_t = 0$ sinon.

On vérifie directement que Φ est un isomorphisme de modules de persistance.

Étant donné un module de persistance \mathbb{V} , on peut définir, pour tout $\epsilon \in \mathbb{R}^+$, l'endomorphisme de \mathbb{V} , $1_{\mathbb{V}}^\epsilon$, donné par la structure de module de persistance :

$$1_{\mathbb{V}}^\epsilon = (v_t^{t+\epsilon} : V_t \rightarrow V_{t+\epsilon})_{t \in \mathbb{R}}.$$

C'est un endomorphisme de degré ϵ , que l'on appelle morphisme de décalage.

Définition 1.3.3. On dit que deux modules de persistance \mathbb{U} et \mathbb{V} sont ϵ -entrelacés s'il existe deux morphismes de degré ϵ , $\Phi : \mathbb{U} \rightarrow \mathbb{V}$ et $\Psi : \mathbb{V} \rightarrow \mathbb{U}$, tels que $\Psi \circ \Phi = 1_{2\epsilon}^{\mathbb{U}}$ et $\Phi \circ \Psi = 1_{2\epsilon}^{\mathbb{V}}$.

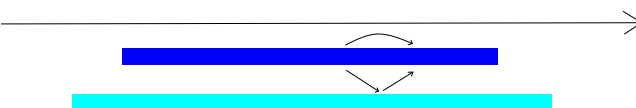
Cette propriété se traduit par les deux diagrammes commutatifs suivants, en tout $t \in \mathbb{R}$:

$$\begin{array}{ccc} U_t & \xrightarrow{u_t^{t+\epsilon}} & U_{t+2\epsilon} \\ & \searrow \phi_t & \nearrow \psi_{t+\epsilon} \\ & V_{t+\epsilon} & \end{array} \quad \text{et} \quad \begin{array}{ccc} & U_{t+\epsilon} & \\ \nearrow \psi_t & & \searrow \phi_{t+\epsilon} \\ V_t & \xrightarrow{v_t^{t+\epsilon}} & V_{t+2\epsilon} \end{array},$$

sans oublier la propriété de morphisme de Φ et Ψ :

$$\begin{array}{ccc} U_s & \xrightarrow{u_s^t} & U_t \\ \downarrow \phi_s & & \downarrow \phi_t \\ V_{s+\epsilon} & \xrightarrow{v_{s+\epsilon}^{t+\epsilon}} & V_{t+\epsilon} \end{array} \quad \text{et} \quad \begin{array}{ccc} U_{s+\epsilon} & \xrightarrow{u_{s+\epsilon}^{t+\epsilon}} & U_{t+\epsilon} \\ \uparrow \psi_s & & \uparrow \psi_t \\ V_s & \xrightarrow{v_s^t} & V_t \end{array}$$

Exemple 1.3.4. Deux modules intervalles $\mathbb{I}([a, b])$ et $\mathbb{I}([c, d])$ sont ϵ -entrelacés si

$$\begin{cases} [a + \epsilon, b - \epsilon] \subset [c, d] \\ \text{et } [c + \epsilon, d - \epsilon] \subset [a, b]. \end{cases}$$


En particulier, un module intervalle est ϵ -entrelacé avec le module trivial si son diamètre est inférieur à ϵ .

Exemple 1.3.5. Reprenons l'exemple 1.2.5 : $(K_t)_{t \in \mathbb{R}}$ et $(L_t)_{t \in \mathbb{R}}$ sont deux familles croissantes de sous-ensembles d'un même espace topologique telles qu'il existe $\epsilon \geq 0, \forall t \in \mathbb{R}$,

$$K_t \subset L_{t+\epsilon} \subset K_{t+2\epsilon}.$$

Ces inclusions induisent les diagrammes :

$$\begin{array}{ccc} H_*(K_t) & \xrightarrow{\quad} & H_*(K_{t+2\epsilon}) \\ & \searrow \quad \nearrow & \\ & H_*(L_{t+\epsilon}) & \end{array} \quad \text{et} \quad \begin{array}{ccc} H_*(L_t) & \xrightarrow{\quad} & H_*(L_{t+2\epsilon}) \\ & \searrow \quad \nearrow & \\ & H_*(K_{t+\epsilon}) & \end{array}$$

Nous avons déjà vu que ces applications définissent bien un ϵ -morphisme entre \mathbb{U} et \mathbb{V} . Nous venons de montrer que c'est même un ϵ -entrelacement.

Définition 1.3.6. Étant donnés deux modules de persistance \mathbb{U} et \mathbb{V} , on définit la distance d'entrelacement (pour l'anglais *interleaving*) :

$$d_i(\mathbb{U}, \mathbb{V}) = \inf \{ \epsilon \geq 0 \text{ tel que } \mathbb{U} \text{ et } \mathbb{V} \text{ soient } \epsilon\text{-entrelacés} \},$$

avec pour convention $\inf \emptyset = +\infty$.

On montre que la distance d'entrelacement est une pseudo-distance (à valeur dans $\mathbb{R}^+ \cup \{\infty\}$) :

Lemme 1.3.7. Soient \mathbb{U}, \mathbb{V} et \mathbb{W} trois modules de persistance, (Φ_1, Ψ_1) un ϵ_1 -entrelacement entre \mathbb{U} et \mathbb{V} , et (Φ_2, Ψ_2) un ϵ_2 -entrelacement entre \mathbb{V} et \mathbb{W} . Alors $(\Phi_2\Phi_1, \Psi_1\Psi_2)$ est un $(\epsilon_1 + \epsilon_2)$ -entrelacement entre \mathbb{U} et \mathbb{W} .

Démonstration. Déjà, on vérifie directement que $\Phi_2\Phi_1$ et $\Psi_1\Psi_2$ sont bien des morphismes de modules de persistance. Ensuite, on calcule :

$$\Psi_1\Psi_2\Phi_2\Phi_1 = \Psi_1 1_{2\epsilon_2}^{\mathbb{V}} \Phi_1.$$

Or, puisque Φ_1 est un ϵ_1 -morphisme, la propriété de commutativité donne $1_{2\epsilon_2}^{\mathbb{V}} \Phi_1 = \Phi_1 1_{2\epsilon_2}^{\mathbb{U}}$. On obtient donc $\Psi_1\Psi_2\Phi_2\Phi_1 = \Psi_1\Phi_1 1_{2\epsilon_2}^{\mathbb{U}} = 1_{2\epsilon_1}^{\mathbb{U}} 1_{2\epsilon_2}^{\mathbb{U}} = 1_{2\epsilon_1+2\epsilon_2}^{\mathbb{U}}$.

On montre de même que $\Phi_2\Phi_1\Psi_1\Psi_2 = 1_{2\epsilon_1+2\epsilon_2}^{\mathbb{W}}$. □

Corollaire 1.3.8. La distance d'entrelacement vérifie l'inégalité triangulaire.

Remarquons toutefois que deux modules peuvent être 0-entrelacés sans pour autant être isomorphes. C'est le cas par exemple de $\mathbb{I}([a, b])$ et $\mathbb{I}([a, b])$.

1.4 Décomposition des modules de persistance

1.4.1 Définitions

Nous reprenons maintenant $T = \mathbb{N}$ ou \mathbb{R} .

Définition 1.4.1. La somme de deux modules de persistance \mathbb{U} et \mathbb{V} est le module de persistance $\mathbb{W} = \mathbb{U} \oplus \mathbb{V}$ où

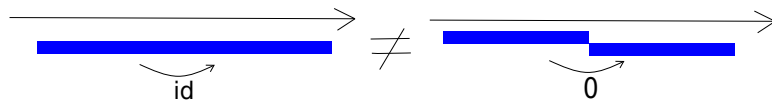
- W_t est la somme directe d'espaces vectoriels $U_t \oplus V_t$
- w_s^t est aussi la somme directe $u_s^t \oplus v_s^t$.

On dit qu'un module de persistance \mathbb{W} est indécomposable si les seules décompositions $\mathbb{W} = \mathbb{U} \oplus \mathbb{V}$ sont les décompositions triviales $\mathbb{W} \oplus 0$ et $0 \oplus \mathbb{W}$.

Les modules intervalles jouent un rôle particulier dans la décomposition des modules de persistance :

Proposition 1.4.2. Les modules intervalles sont indécomposables.

Démonstration. On écrit $\mathbb{I} = \mathbb{U} \oplus \mathbb{V}$ avec \mathbb{I} un module intervalle. On vérifie immédiatement que l'application de projection $\mathbb{U} \oplus \mathbb{V} \rightarrow \mathbb{U}$ définit un 0-endomorphisme du module \mathbb{U} . Or les endomorphismes de \mathbb{I} forment une k -algèbre isomorphe à k (proposition 1.2.6). La projection est idempotente, et les seuls idempotents de k sont 0 ou 1. On déduit que $\mathbb{U} = 0$ ou $\mathbb{V} = 0$. \square



On dit qu'un module de persistance \mathbb{V} se décompose en modules-intervalles si l'on a un isomorphisme entre \mathbb{V} et une somme de modules-intervalles :

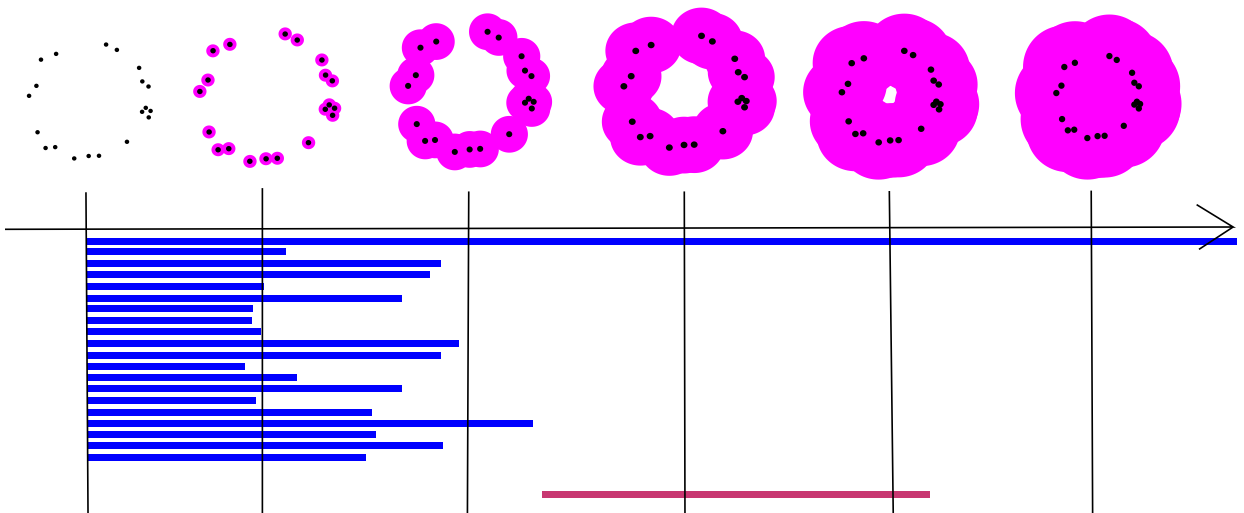
$$\mathbb{V} \simeq \bigoplus_{l \in L} \mathbb{I}(a_l^*, b_l^*),$$

où la notation c^* signifie c^+ ou c^- en fonction de l'appartenance de c à l'intervalle.

On peut trouver dans [1] la preuve du résultat suivant, qui est une conséquence du théorème de Krull–Remak–Schmidt–Azumaya :

Théorème 1.4.3. Une décomposition (éventuellement infinie) en modules-intervalles est unique.

Topologiquement parlant, si le module de persistance \mathbb{V} représente l'évolution des l'homologie d'un espace (comme dans l'exemple introductif), alors on interprète les sommandes de la décomposition $\mathbb{V} = \bigoplus_{l \in L} \mathbb{I}(a_l^*, b_l^*)$ comme des caractéristiques topologiques présentes dans les intervalles $[a_l^*, b_l^*]$.



La décomposition de $H_0(X^t)$ est représentée en bleu, et celle de $H_1(X^t)$ en rouge.

Si l'on connaît la décomposition de \mathbb{V} en modules-intervalle, on peut ainsi résumer son homologie persistante par une famille d'intervalles.

Nous terminons cette partie par le résultat suivant :

Lemme 1.4.4. Soient $\mathbb{U}_1, \mathbb{U}_2, \mathbb{V}_1, \mathbb{V}_2$ des modules de persistance. On a :

$$d_i(\mathbb{U}_1 \oplus \mathbb{U}_2, \mathbb{V}_1 \oplus \mathbb{V}_2) \leq \max(d_i(\mathbb{U}_1, \mathbb{V}_1), d_i(\mathbb{U}_2, \mathbb{V}_2)).$$

Démonstration. Donnons-nous

- $\epsilon_1 \geq d_i(\mathbb{U}_1, \mathbb{V}_1)$ et (Φ_1, Ψ_1) un ϵ_1 -entrelacement entre \mathbb{U}_1 et \mathbb{V}_1 ,
- $\epsilon_2 \geq d_i(\mathbb{U}_2, \mathbb{V}_2)$ et (Φ_2, Ψ_2) un ϵ_2 -entrelacement entre \mathbb{U}_2 et \mathbb{V}_2 .

Il suffit de montrer que la somme $(\Phi_1 \oplus \Phi_2, \Psi_1 \oplus \Psi_2)$ est un $(\epsilon_1 + \epsilon_2)$ -entrelacement entre $\mathbb{U}_1 \oplus \mathbb{U}_2$ et $\mathbb{V}_1 \oplus \mathbb{V}_2$. Or ceci est clair puisque la somme d'applications linéaires est définie coordonnées par coordonnées. \square

Dans les trois parties suivantes, nous présentons des cas dans lesquels l'existence d'une telle décomposition est connue.

1.4.2 Décomposition des modules de type fini sur \mathbb{N}

Cette partie suit l'exposition de [2].

On suppose que $T = \mathbb{N}$. Dans ce contexte, les intervalles sont $[a, b] = \{a, a + 1, \dots, b\}$ et $[a, b[= \{a, a + 1, \dots, b - 1\}$.

Nous allons obtenir la décomposition d'un module de persistance \mathbb{V} sur \mathbb{N} en le regardant comme un module sur un anneau principal. Pour obtenir une bonne correspondance entre ces deux structures, il sera même plus commode de le voir comme un module gradué.

Notons $k[X]$ l'anneau des polynômes sur le corps k . En tant que groupe additif, il s'écrit comme la somme directe $k[X] = \bigoplus_{i \in \mathbb{N}} k[X]_i$, avec $k[X]_i$ l'ensemble des monômes de degré i . Puisque la multiplication vérifie

$$k[X]_i \cdot k[X]_j \subset k[X]_{i+j},$$

on dit que l'on a une graduation de $k[X]$.

Soit $\mathbb{V} = \{V_i, v_i^j\}$ un module de persistance. Nous définissons l'espace vectoriel :

$$M = \bigoplus_{i \in \mathbb{N}} V_i.$$

Nous allons construire une structure de $k[X]$ -module gradué sur M . Pour tout $i, n \in \mathbb{N}$, identifions l'application linéaire
$$\begin{array}{ccc} V_n & \rightarrow & V_{n+i} \\ x & \mapsto & v_n^{n+i}(x) \end{array}$$
 par la multiplication extérieure par X^i . Coordonnées par coordonnées, on obtient une application :

$$\begin{array}{ccc} k[X] \times M & \longrightarrow & M \\ \sum \lambda_i X^i, x & \longrightarrow & \sum \lambda_i X^i \cdot x \end{array}$$

Plus explicitement, le produit de X^i par un élément $(x_n)_{n \in \mathbb{N}} \in M$ est $(v_{n-i}^n(x_{n-i}))_{n \in \mathbb{N}}$, où les i premiers termes sont des zéros.

On a ainsi muni M d'une structure de module sur $k[X]$. De plus, la décomposition $M = \bigoplus_{i \in \mathbb{N}} V_i$ vérifie :

$$k[X]_i \cdot V_j \subset V_{i+j}.$$

On dit que M est muni d'une structure de module gradué.

Supposons maintenant que M soit un module finiment engendré. Pour comprendre où intervient la structure de module gradué dans notre problème, commençons par utiliser le théorème de décomposition d'un module de type fini sur un anneau principal : il existe un entier l et une suite de polynômes $(P_i)_{i \leq m}$, avec $P_i | P_{i+1}$, tels que M soit isomorphe (en tant que module) à

$$M = k[X]^l \bigoplus_{i \leq m} k[X] / P_i$$

Nous aimerions identifier les composantes $k[X] / P_i$ de la décomposition de M avec des modules-intervalle \mathbb{I} de la décomposition de \mathbb{V} . Déjà, un intervalle associé à $k[X] / P_i$ devrait contenir $\dim k[X] / P_i = \deg(P_i)$ éléments. Il nous manque encore une information pour identifier l'intervalle en question. Pour cela, nous allons utiliser la structure graduée.

Définition 1.4.5. Un isomorphisme de modules gradués entre deux modules gradués $M = \bigoplus_{i \in \mathbb{N}} V_i$ et $N = \bigoplus_{i \in \mathbb{N}} W_i$ est un isomorphisme de modules $\Phi : M \rightarrow N$ tel que pour tout $i \in \mathbb{N}$, $\Phi(V_i) = W_i$.

Si L est un sous-ensemble de M , on note $\Sigma^i L = X^i \cdot L$, l'opérateur de décalage dans la graduation. On a alors un théorème de classification plus précis : il existe un entier l , des entiers L_1, \dots, L_l , une suite d'entiers $(d_i)_{i \leq m}$, et des entiers D_1, \dots, D_d tels que M soit isomorphe (en tant que module gradué) à

$$M = \bigoplus_{1 \leq i \leq l} \Sigma^{L_i} k[X] \bigoplus_{1 \leq i \leq m} \Sigma^{D_i} (k[X] / X^{d_i}),$$

où la graduation de $\Sigma^{D_i} (k[X] / X^{d_i})$ est celle héritée de $\Sigma^{D_i} (k[X])$.

On voit ici apparaître les modules-intervalles de la décomposition de \mathbb{V} :

- une composante $\Sigma^{D_i} (k[X] / X^{d_i})$ correspond à $\mathbb{I}([D_i, D_i + d_i[)$,
- et $\Sigma^{L_i} k[X]$ correspond à $\mathbb{I}([L_i, +\infty[)$.

Les trois résultats suivants permettent de justifier cela.

Lemme 1.4.6. Soient \mathbb{U} et \mathbb{V} deux modules de persistance, $M_{\mathbb{U}}$ et $M_{\mathbb{V}}$ leurs modules gradués associés. Si $M_{\mathbb{U}}$ et $M_{\mathbb{V}}$ sont isomorphes, alors \mathbb{U} et \mathbb{V} sont isomorphes.

Démonstration. On note $\mathbb{U} = (U_i, u_i^j)$, $\mathbb{V} = (V_i, v_i^j)$, $M_{\mathbb{U}} = \bigoplus_{i \in \mathbb{N}} U_i$ et $M_{\mathbb{V}} = \bigoplus_{i \in \mathbb{N}} V_i$. Soit $\Phi : M_{\mathbb{U}} \rightarrow M_{\mathbb{V}}$ l'isomorphisme de modules gradués en question. Puisque Φ respecte les graduations, il induit des isomorphismes $\phi_i : U_i \rightarrow V_i$. On veut montrer que cette famille forme un isomorphisme de modules de persistance. Autrement dit, on veut que le diagramme suivant commute :

$$\begin{array}{ccc} V_i & \xrightarrow{v_i^j} & V_j \\ \downarrow \phi_i & & \downarrow \phi_j \\ W_i & \xrightarrow{w_i^j} & W_j \end{array}$$

Soient $\mu_i : U_i \rightarrow U_i \subset M_{\mathbb{U}}$ les applications qui envoient chaque élément sur la i -ème coordonnée. Si $x \in U_i \subset M_{\mathbb{U}}$, alors $\mu_i^{-1}(x)$ est bien défini. De même, on définit $\nu_i : V_i \rightarrow V_i \subset M_{\mathbb{V}}$ et ν_i^{-1} .

D'autre part, puisque Φ est un morphisme de modules, on a, pour tout polynôme $P \in k[X]$, le diagramme commutatif :

$$\begin{array}{ccc} M_{\mathbb{U}} & \xrightarrow{P \times} & M_{\mathbb{U}} \\ \downarrow \Phi & & \downarrow \Phi \\ M_{\mathbb{V}} & \xrightarrow{P \times} & M_{\mathbb{V}} \end{array}$$

où l'on note $P \times$ la multiplication extérieure par P . Il suffit de transposer ce diagramme aux modules de persistance.

Considérons pour cela le diagramme suivant :

$$\begin{array}{ccccc}
 M_{\mathbb{U}} & \xrightarrow{X^{j-i} \times} & M_{\mathbb{U}} & & \\
 \downarrow \Phi & \swarrow \mu_i & U_i \xrightarrow{u_i^j} U_j & \swarrow \mu_j^{-1} & \downarrow \Phi \\
 & & \downarrow \phi_i & & \downarrow \phi_j \\
 & & V_i \xrightarrow{v_i^j} V_j & & \\
 & \swarrow \nu_i & & \swarrow \nu_j^{-1} & \\
 M_{\mathbb{V}} & \xrightarrow{X^{j-i} \times} & M_{\mathbb{V}} & &
 \end{array}$$

Les trapèzes supérieurs et inférieurs commutent par définition du module gradué induit par un module de persistance. Les trapèzes à gauche et à droite commutent par définition des ϕ_i . Enfin, le carré extérieur commute par isomorphisme.

On en conclut que le carré intérieur commute. Nous avons donc bien construit un isomorphisme de modules de persistance. \square

Lemme 1.4.7. Soient $a < b \in \mathbb{N}$, et $\mathbb{I}([a, b])$ le module-intervalle.

- Le module gradué M précédemment construit se décompose (en tant que module gradué) en $M = \Sigma^a k[X]/X^{b-a}$.
- Réciproquement, si \mathbb{V} est un module de persistance, et que le module gradué associé se décompose en $M = \Sigma^a k[X]/X^{b-a}$, alors \mathbb{V} est isomorphe au module intervalle $\mathbb{I}([a, b])$.

Démonstration. Premier point : pour commencer, nous étudions la graduation de $k[X]/X^{b-a}$.

La classe d'équivalence de $x \in k[X]$ dans le quotient $k[X]/X^{b-a}$ s'écrit $x + X^{b-a} \cdot k[X]$. De même, l'ensemble des monômes de degré i est représenté dans $k[X]/X^{b-a}$ par $k[X]_n + X^{b-a} \cdot k[X]$. On obtient ainsi une graduation de $k[X]/X^{b-a}$ à partir de la graduation de $k[X]$:

$$k[X]/X^{b-a} = \bigoplus_{0 \leq i < b-a} (k[X]_i + X^{b-a} \cdot k[X]).$$

De même, on obtient une graduation de $\Sigma^a k[X]/X^{b-a}$:

$$\Sigma^a k[X]/X^{b-a} = \bigoplus_{a \leq i < b} (k[X]_i + X^{b-a} \cdot k[X]).$$

Soit I le groupe additif défini par $I = \bigoplus_{i \in \mathbb{N}} W_i$, où $W_i = k$ si $a \leq i < b$, et $W_i = 0$ sinon. $\Sigma^a k[X]/X^{b-a}$ et I sont isomorphes en tant que groupe. Par transfert de structure, on obtient un isomorphisme de modules gradués entre $\Sigma^a k[X]/X^{b-a}$ et $I = \bigoplus_{a \leq i < b} k$.

Maintenant, remarquons que M est défini par $M = \bigoplus_{a \leq i < b} k$. Par définition de la structure de module gradué de M , on vérifie directement que M et I sont isomorphes.

Deuxième point : on suppose maintenant que M est isomorphe à $k[X]/X^{b-a}$. D'après le premier point, $k[X]/X^{b-a}$ est isomorphe au module gradué induit par le module-intervalle $\mathbb{I}([a, b])$. Par le lemme 1.4.6, on obtient que M et $\mathbb{I}([a, b])$ sont isomorphes. \square

On obtient alors la correspondance suivante :

Proposition 1.4.8. Soit \mathbb{V} un module de persistance.

- Si $\mathbb{V} = \bigoplus_{1 \leq l \leq L} \mathbb{I}([a_l, b_l])$ se décompose en modules intervalles, alors son module gradué associé se décompose en

$$M = \bigoplus_{1 \leq l \leq L} \Sigma^{a_l} k[X]/X^{b_l - a_l},$$

avec pour convention $\Sigma^{a_l} k[X]/X^{+\infty - a_l} = \Sigma^{a_l} k[X]$.

• Si le module gradué associé à \mathbb{V} se décompose en $M = \bigoplus_{1 \leq i \leq l} \Sigma^{L_i} k[X] \oplus \bigoplus_{1 \leq i \leq m} \Sigma^{D_i} k[X]/X^{d_i}$, alors on a la décomposition

$$\mathbb{V} = \bigoplus_{1 \leq i \leq l} \mathbb{I}([L_i, +\infty[) \oplus \bigoplus_{1 \leq i \leq m} \mathbb{I}([D_i, D_i + d_i])$$

Démonstration. Premier point : Remarquons que si \mathbb{U} et \mathbb{V} sont deux modules de persistance de modules gradués associés $M = \bigoplus U_i$ et $N = \bigoplus V_i$, alors le module associé à $\mathbb{U} \oplus \mathbb{V}$ est $\bigoplus U_i \oplus V_i$. Il suffit alors d'appliquer le lemme 1.4.7.

Deuxième point : c'est une conséquence du lemme 1.4.6 et du point précédent. \square

Remarque 1.4.9. Si le module gradué associé à un module de persistance est finiment engendré, on vient de montrer que ce dernier se décompose comme somme de modules-intervalle. On verra que c'est le cas par exemple des modules construits sur les voisinages tubulaires d'ensembles finis de points, comme dans la partie 1.1.

Pour terminer cette partie, nous citons un résultat qui permet en particulier de retrouver le fait que la décomposition d'un module de persistance est unique. On note r_a^b le rang de l'application $v_a^b : V_a \rightarrow V_b$.

Proposition 1.4.10. Si $\mathbb{V} = \bigoplus_{1 \leq l \leq L} \mathbb{I}([a_l, b_l])$ est un module de persistance sur \mathbb{N} décomposé en modules-intervalle, alors :

- pour $a < b \in N$, le nombre d'intervalles $[a, b[$ dans sa décomposition est

$$r_a^{b-1} - r_a^b + r_{a-1}^b - r_{a-1}^{b-1}$$

- pour $a \in N$, le nombre d'intervalles $[a, +\infty[$ dans sa décomposition est

$$\lim_{b \rightarrow +\infty} r_a^b - \lim_{b \rightarrow +\infty} r_{a-1}^b$$

1.4.3 Décomposition des modules modérés sur \mathbb{R}^+

Nous nous intéressons maintenant aux modules de persistance \mathbb{V} sur \mathbb{R}^+ . En suivant [3], on définit ses valeurs critiques comme les réels $t \in \mathbb{R}^+$ tels que pour tout $\epsilon > 0$ assez petit, l'application $v_{t-\epsilon}^{t+\epsilon}$ n'est pas un isomorphisme.

Définition 1.4.11. Le module de persistance \mathbb{V} est dit modéré (de l'anglais *tame*) s'il n'a qu'un nombre fini de valeurs critiques, et que chaque $V_t, t \in \mathbb{R}^+$ est de dimension finie.

Proposition 1.4.12. Si \mathbb{V} ne contient pas de valeur critique dans l'intervalle $[a, b]$, alors toutes les applications v_s^t , avec $a \leq s < t \leq b$, sont des isomorphismes.

Démonstration. Par l'absurde, on suppose qu'une telle application v_s^t ne soit pas un isomorphisme. Si on note $u = \frac{s+t}{2}$, alors au moins l'une des applications v_s^u et v_u^t n'est pas un isomorphisme (car $v_u^t \circ v_s^u = v_s^t$). En répétant cette opération, on converge vers une valeur critique de \mathbb{V} dans $[a, b]$. \square

Nous allons maintenant établir une correspondance entre les modules de persistance modérés sur \mathbb{R}^+ et les modules de persistance sur \mathbb{N} . Pour cela, définissons les fonctions grille comme les fonctions $T : \mathbb{N} \rightarrow \mathbb{R}$ croissantes et sans points d'accumulation.

Soit maintenant \mathbb{V} un module de persistance modéré et $C = \{c_1, \dots, c_n\}$ l'ensemble de ses valeurs critiques. Soit $D = \{d_0, \dots, d_{n+1}\} \subset \mathbb{R}^+$ une suite de réels entrelacée avec C :

$$d_0 < c_1 < d_1 < \dots < c_n < d_{n+1}.$$

On choisit T une fonction grille qui contient C et D dans son image. On définit alors le **module de persistance discrétisé** \mathbb{U} sur \mathbb{N} par les espaces $U_i = V_{T(i)}$, et les applications entre ces espaces données par la structure de module de persistance de \mathbb{V} .

La proposition précédente montre que \mathbb{V} est entièrement déterminé par le module \mathbb{U} , à isomorphisme près. De plus, on peut montrer que la décomposition en module gradué de $M_{\mathbb{U}}$ se transpose en une décomposition en modules-intervalles de \mathbb{V} .

1.4.4 Décomposition des modules finis partout sur \mathbb{R}

Dans [4], l'auteur généralise les résultats précédents :

Théorème 1.4.13. Soit \mathbb{V} un module de persistance sur \mathbb{R} tel que chaque espace V_t soit de dimension finie. Alors \mathbb{V} se décompose en somme (éventuellement infinie) de modules-intervalles.

Il s'appuie pour cela sur la notion suivante :

Définition 1.4.14. Le module \mathbb{V} satisfait la condition de chaîne descendante sur les images et les noyaux si :

- pour tout réel $t \in \mathbb{R}$ et toute suite décroissante $t \geq s_1 > s_2 > \dots$, la suite

$$V_t \supset v_{s_1}^t \supset v_{s_2}^t \supset \dots$$

est stationnaire, et

- pour tout réel $t \in \mathbb{R}$ et toute suite décroissante $r_1 > r_2 > \dots \geq t$, la suite

$$V_t \supset v_t^{r_1} \supset v_t^{r_2} \supset \dots$$

est stationnaire.

Théorème 1.4.15. Tout module de persistance sur \mathbb{R} vérifiant la condition de chaîne descendante sur les images et les noyaux se décompose en somme de modules-intervalle.

1.5 Diagrammes de persistance

1.5.1 Définition

Le diagramme de persistance est l'objet phare de l'analyse topologique des données. C'est une façon de représenter la décomposition en modules-intervalle d'un module de persistance.

Considérons un module décomposé :

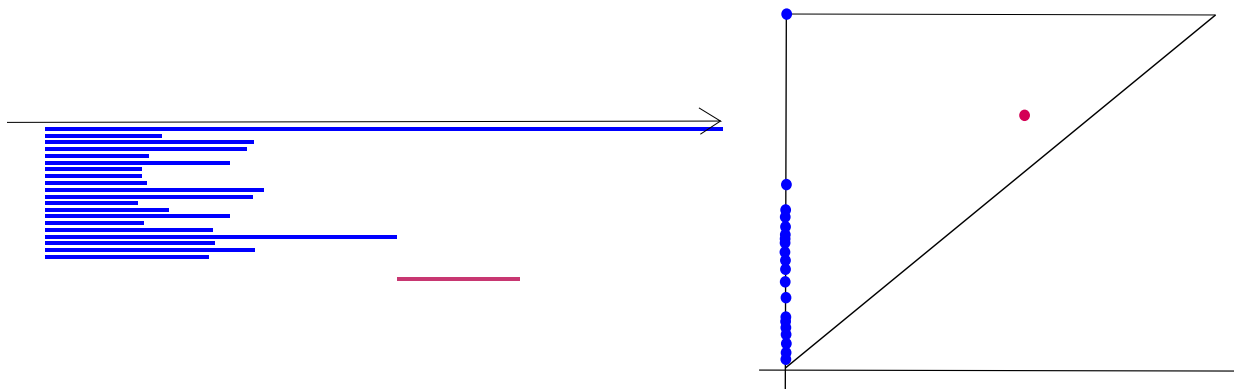
$$\mathbb{V} = \bigoplus_{l \in L} \mathbb{I}(a_l^*, b_l^*),$$

où l'étoile en indice signifie que l'intervalle est ouvert ou fermé de ce côté l'intervalle.

Définition 1.5.1. Le diagramme de persistance de \mathbb{V} est le multi-ensemble

$$\text{dgm}(\mathbb{V}) := \{(a_l, b_l), l \in L\} \setminus \Delta \subset \bar{\mathbb{R}}^2,$$

où les points sont comptés avec multiplicité. On a retiré du diagramme les points (a_l, b_l) de la diagonale $\Delta = \{(x, x), x \in \mathbb{R}\}$.



Le théorème 1.4.3 assure que $\text{dgm}(\mathbb{V})$ ne dépend pas de la décomposition de \mathbb{V} .

La distance bottleneck. Étant donnés deux modules décomposables \mathbb{U} et \mathbb{V} , on veut pouvoir les comparer à travers leurs diagrammes $\text{dgm}(\mathbb{U}), \text{dgm}(\mathbb{V})$. Deux diagrammes sont proches si on peut appareiller leurs points sans trop les déplacer. On formalise cette idée ainsi :

- Un appariement partiel entre deux multi-ensembles $A, B \subset \mathbb{R}$ est un sous-multi-ensemble $M \subset A \times B$ tel que :

- pour tout $a \in A$, il y a au plus un $b \in B$ tel que $(a, b) \in M$,
- pour tout $b \in B$, il y a au plus un $a \in A$ tel que $(a, b) \in M$.

De tels points sont dit appariés. On considère que les autres points sont appariés avec la diagonale Δ .

- Pour $\delta \geq 0$, on dit que l'appariement M est un δ -appariement si :

- si $(a, b) \in M$, alors $d_\infty(a, b) \leq \delta$,
- si $a \in A$ n'est pas apparié, alors $d_\infty(a, \Delta) \leq \delta$,
- si $b \in B$ n'est pas apparié, alors $d_\infty(b, \Delta) \leq \delta$.

Définition 1.5.2. La distance bottleneck entre \mathbb{U} et \mathbb{V} est définie par :

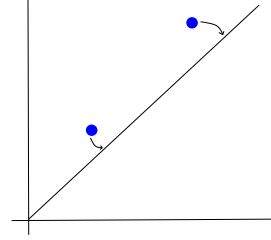
$$d_b(\mathbb{U}, \mathbb{V}) = \inf\{\delta \geq 0 \text{ tel qu'il existe un } \delta\text{-appariement entre } \text{dgm}(\mathbb{U}) \text{ et } \text{dgm}(\mathbb{V})\}$$

Proposition 1.5.3. Si $U = \mathbb{I}(a^*, b^*)$ et $V = \mathbb{I}(c^*, d^*)$ sont deux modules-intervalle, alors

$$d_b(\mathbb{U}, \mathbb{V}) \leq d_\infty((a, b), (c, d)).$$

Démonstration. Il suffit d'utiliser l'appariement $M = ((a, b), (c, d))$. □

La figure à droite montre que l'inégalité précédente n'est pas toujours une égalité. Il s'agit de deux points, pour lesquels il est moins coûteux de les appairer avec la diagonale plutôt qu'entre eux.



On peut trouver les résultats suivants dans [1] :

Proposition 1.5.4. Si A , B et C sont trois multi-ensembles, alors

$$d_b(A, C) \leq d_b(A, B) + d_b(A, C).$$

Proposition 1.5.5. Soient A et B sont deux multi-ensembles. Si $d_b(A, B) = \delta$, alors il existe un δ -entrelacement entre eux.

1.5.2 Stabilité

Considérons deux modules de persistance \mathbb{U} et \mathbb{V} décomposables en modules-intervalle. Nous avons défini jusqu'à présent deux notions de distance entre ces modules :

- la distance d'entrelacement d_i , qui compare leurs structures de modules,
- et la distance bottleneck, d_b , qui compare leurs diagrammes de persistance.

Ces deux distances présentent des qualités propres : la première s'exprime au niveau algébrique, et la deuxième se manipule algorithmiquement. En fait, il se trouve que ces deux notions coïncident, comme nous le verrons dans la partie 1.6 : $d_i = d_b$.

Avec peu d'efforts, nous pouvons déjà démontrer la moitié de ce résultat. C'est ce que l'on appelle le théorème d'isométrie :

Théorème 1.5.6. Si \mathbb{U} et \mathbb{V} sont deux modules de persistance (sur \mathbb{R}) décomposables en modules-intervalle, alors

$$d_i(\mathbb{U}, \mathbb{V}) \leq d_b(\mathbb{U}, \mathbb{V}).$$

Démonstration. Notons $\mathbb{U} = \bigoplus_{l \in L} \mathbb{I}(a_l^*, b_l^*)$ et $\mathbb{V} = \bigoplus_{k \in K} \mathbb{I}(c_k^*, d_k^*)$.

Soit $\delta \geq d_b(\mathbb{U}, \mathbb{V})$. Nous allons montrer que $d_i(\mathbb{U}, \mathbb{V}) \leq \delta$. Soit $M \subset \text{dgm}(\mathbb{U}) \times \text{dgm}(\mathbb{V})$ un δ -appariement. M définit :

- Des points appariés $((a_l, b_l), (c_k, d_k))$ tels que $d_\infty((a_l, b_l), (c_k, d_k)) \leq \delta$. La proposition 1.5.3 nous assure alors que $d_b(\mathbb{I}(a_l^*, b_l^*), \mathbb{I}(c_k^*, d_k^*)) \leq \delta$.
- Des points non-appariés (a_l, b_l) ou (c_k, d_k) qui sont δ -proches de la diagonale. De la même façon, ces modules-intervalle sont δ -entrelacés avec le module trivial.

Maintenant, il suffit de sommer tous ces modules intervalles et d'appliquer le lemme 1.4.4 pour obtenir :

$$d_i\left(\bigoplus_{l \in L} \mathbb{I}(a_l^*, b_l^*), \bigoplus_{k \in K} \mathbb{I}(c_k^*, d_k^*)\right) \leq \delta.$$

□

1.6 Le théorème d'isométrie

1.6.1 Cas des modules modérés

L'objet de cette partie est de présenter le théorème d'isométrie :

Théorème 1.6.1. Si \mathbb{U} et \mathbb{V} sont deux modules de persistance (sur \mathbb{R}) modérés, alors

$$d_i(\mathbb{U}, \mathbb{V}) = d_b(\mathbb{U}, \mathbb{V}).$$

Nous avons déjà démontré l'inégalité $d_i(\mathbb{U}, \mathbb{V}) \leq d_b(\mathbb{U}, \mathbb{V})$ au théorème 1.5.6. Expliquons maintenant pourquoi l'inégalité inverse demande plus de travail. Si l'on dispose d'un ϵ -entrelacement (Φ, Ψ) entre \mathbb{U} et \mathbb{V} , il se peut que ces applications mélangent les modules intervalles. Autrement dit, les isomorphismes peuvent ne pas être diagonaux dans les bases données par la décomposition en modules-intervalles.

Or un appariement entre les diagrammes $\text{dgm}(\mathbb{U})$ et $\text{dgm}(\mathbb{V})$ demande de définir des morphismes qui envoient un unique intervalle sur un autre. Le problème est donc de construire un entrelacement qui respecterait cette décomposition. Pour ce faire, nous pouvons utiliser le lemme d'interpolation :

Lemme 1.6.2. Si les modules \mathbb{U} et \mathbb{V} sont ϵ -entrelacés, alors il existe une famille de modules de persistance, $\{\mathbb{U}_x, x \in [0, \epsilon]\}$, tels que $\mathbb{U}_0 = \mathbb{U}$, $\mathbb{U}_\epsilon = \mathbb{V}$, et

$$\forall x, y \in [0, \epsilon], d_i(\mathbb{U}_x, \mathbb{U}_y) \leq |x - y|.$$

On pourra par exemple trouver une preuve de ces résultats dans [5].

1.6.2 Généralisation aux modules q -modérés

Dans cette dernière partie, nous évoquons les résultats de [1], qui généralisent ce que nous avons vu jusque maintenant.

Mesure persistante. La définition que l'on a donné en 1.5.1 des diagrammes de persistance demande au préalable la décomposabilité en modules-intervalle. On peut se débarrasser de cette condition grâce à la notion de mesure persistante. Par la suite, nous noterons r_s^t le rang de l'application $v_s^t : V_s \rightarrow V_t$.

Définition 1.6.3. Soit \mathbb{V} un module de persistance. Sa mesure persistante est la fonction

$$\mu_{\mathbb{V}} : R \longmapsto r_a^b - r_a^c + r_a^d - r_b^d,$$

définie sur les rectangles $R = [a, b] \times [c, d]$ tels que $a < b \leq c < d$.

Intuitivement, cette fonction permet de localiser les points sur le diagramme de persistance de \mathbb{V} . Par exemple, la mesure persistante d'un module intervalle $\mathbb{I}(I)$ est

$$\mu_{\mathbb{I}}(R) = \begin{cases} 1 & \text{si } [b, c] \subset I \subset [a, d], \\ 0 & \text{sinon.} \end{cases}$$

Plus généralement :

Proposition 1.6.4. Soit $\mathbb{V} = \bigoplus_{l \in L} \mathbb{I}(x_l^*, y_l^*)$ décomposé en modules-intervalle, et $R = [a, b] \times [c, d]$. Alors $\mu_{\mathbb{V}}(R)$ est le nombre de points (x_l^*, y_l^*) tels que $[b, c] \subset (x_l^*, y_l^*) \subset [a, d]$.

Définition 1.6.5. Un module de persistance \mathbb{V} est dit q -modéré (de l'anglais *q-tame*) si sa mesure persistante prend des valeurs finies sur tout les rectangles R qui n'intersectent pas la diagonale.

De manière équivalente, \mathbb{V} est q -modéré si tous les morphismes $v_s^t, s < t$, sont de rang fini. Nous verrons par la suite que cette hypothèse est suffisamment faible pour contenir tous les cas que nous étudierons.

Diagramme d'une mesure persistante. Soit A un multi-ensemble de \mathbb{R}^2 . Comme pour une mesure de persistance, on peut définir une fonction sur les rectangles :

$$R \longmapsto \text{card}(A \cap R).$$

La proposition suivante permet de définir le diagramme d'une mesure persistante :

Proposition 1.6.6. Soit $\mu_{\mathbb{U}}$ la mesure persistante d'un module de persistance q -modéré. Alors il existe un unique multi-ensemble localement fini $\text{Dgm}(\mu_{\mathbb{V}}) \subset \{(x, y) \in \mathbb{R}^2, x < y\}$ tel que pour tout rectangle R qui n'intersecte pas la diagonale,

$$\mu_{\mathbb{V}}(R) = \text{card}(\text{Dgm}(\mu_{\mathbb{V}}) \cap R).$$

Ainsi on peut définir, pour deux diagrammes de persistance q -modérés \mathbb{U} et \mathbb{V} leur distance bottleneck

$$d_b(\mathbb{U}, \mathbb{V}) = d_b(\text{Dgm}(\mu_{\mathbb{U}}), \text{Dgm}(\mu_{\mathbb{V}})).$$

Dans ce cadre, on peut énoncer le théorème d'isométrie pour les modules q -modérés :

Théorème 1.6.7. Si \mathbb{U} et \mathbb{V} sont deux modules de persistance q -modérés, alors

$$d_i(\mathbb{U}, \mathbb{V}) = d_b(\mathbb{U}, \mathbb{V}).$$

2 Espaces topologiques filtrés

En pratique, on obtient un module de persistance de deux façon : en filtrant un espace topologique, ou en filtrant un complexe simplicial. On s'occupe ici du premier cas.

2.1 Filtrations

Soit E un espace topologique et $f : E \rightarrow \mathbb{R}$ une fonction continue. Les sous-niveaux $f^t = f^{-1}(]-\infty, t])$ de f forment une famille croissante de sous-espaces topologiques indexée par \mathbb{R} . Pour tout $r \leq s \leq t \in \mathbb{R}$,

$$f^r \subset f^s \subset f^t.$$

Si l'on définit par $V_t = H_*(f^t)$ les groupes d'homologie singulière, et $v_s^t : V_t = H_*(f^s) \rightarrow H_*(f^t)$ les morphismes induits par les inclusions, on obtient un module de persistance

$$\mathbb{V}_f = ((V_t)_{t \in \mathbb{R}}, (v_s^t)_{s \leq t \in \mathbb{R}}).$$

Définition 2.1.1. Si $f : X \rightarrow \mathbb{R}$ est différentiable (avec X un espace euclidien ou une variété différentielle), on appelle point critique de f un point $x \in X$ tel que $\nabla f(x) = 0$. Une valeur critique de f est un réel $c \in \mathbb{R}$ tel que $f^{-1}(\{c\})$ contienne au moins un point critique. Sinon, elle est dite régulière.

Dans ce cas, nous allons voir que la l'évolution de la topologie des sous-niveaux f^t dépend des points critiques de la fonction f .

2.2 Fonctions de Morse

Soit X une variété différentielle, et $f : X \rightarrow \mathbb{R}$ une fonction de classe \mathcal{C}^2 . On dit que f est une fonction de Morse si ses points critiques sont non-dégénérés (c'est à dire que la hessienne $d^2f(x)$ est une forme quadratique non dégénérée).

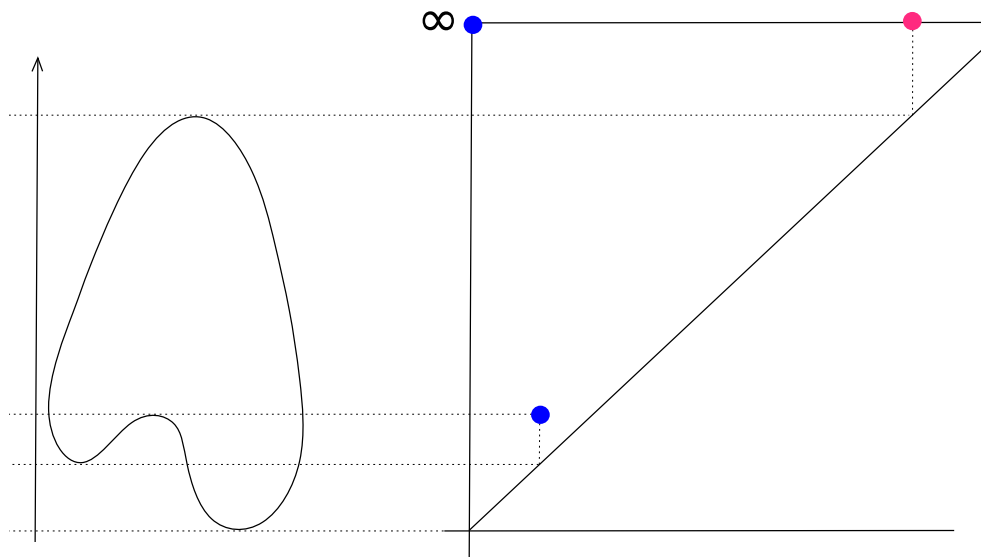
On sait que les points critiques d'une telle fonction sont isolés. Ce sont les points où la topologie des sous-niveaux change.

En dehors de ces points critiques, on peut définir des champs de vecteurs qui rétractent les sous-niveaux de f . De cette façon, on montre dans [6] que :

Théorème 2.2.1. Soit $a, b \in \mathbb{R}$ tels que f n'admette pas de valeur critique dans l'intervalle $[a, b]$. Supposons que $f^{-1}([a, b])$ soit compact. Alors f^b se rétracte par déformation sur f^a .

Soit \mathbb{V}_f le module de persistance engendré par f . Pour tout réel t , on sait que le sous-niveau f^t est une variété topologique à bord, et que son groupe d'homologie singulière est de dimension finie. Si de plus le nombre de points critiques de f est fini (par exemple si X est compacte), on est dans le cadre de la partie 1.4.3 : \mathbb{V}_f est un module modéré, et il se décompose en somme finie de modules-intervalle.

L'exemple suivant est le diagramme de persistance d'un courbe de \mathbb{R}^2 filtrée par la projection sur l'axe vertical.



Si la variété X en question est compacte, alors les sous-niveaux sont constants à partir d'une certaine valeur de t . A ce moment, les groupes d'homologie de f^t valent ceux de X , et les applications v_s^t sont l'identité. En particulier, les points du diagramme qui persistent jusqu'à l'infini sont en correspondance avec les nombres de Betti de X .

2.3 Fonctions distance

Les fonction distance sont des filtrations naturelles lorsque l'on s'intéresse à un compact. Nous pouvons retrouver des propriétés analogues aux fonctions de Morse.

2.3.1 Axe médian

Soit K un compact de l'espace euclidien \mathbb{R}^d . Pour tout $x \in \mathbb{R}^d$, l'infimum $\inf_{k \in K} \|x - k\|$ est atteint, et définit la fonction distance à K :

$$d_K : x \in \mathbb{R}^d \mapsto \inf_{k \in K} \|x - k\|.$$

On appelle ensemble des projections de x sur K l'ensemble :

$$\Gamma(x) = \operatorname{argmin}_{k \in K} \|x - k\|.$$

Soient $x, y \in \mathbb{R}^d$, et $k_x \in \Gamma(x)$ une projection de x sur K . On a :

$$d_K(y) - d_K(x) \leq \|y - k_x\| - \|x - k_x\| \leq \|x - y\|.$$

Par symétrie, on obtient que la fonction d_K est 1-lipschitzienne.

Les sous-niveaux de d_K sont les voisinages tubulaires K^t . Comme pour les fonctions régulières, on voudrait étudier l'évolution des sous-niveaux de d_K grâce à ses valeurs critiques. Malheureusement, cette fonction n'est pas différentiable partout.

Par exemple, la figure à droite représente la fonction distance à une ellipse. On y observe sa non-différentiabilité au centre du domaine.

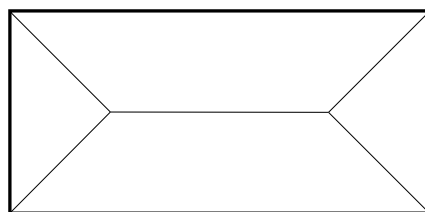
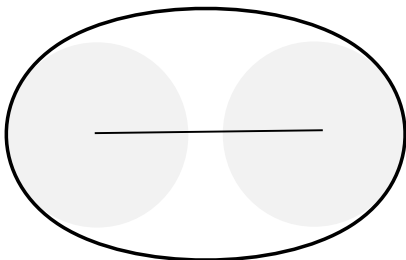


On peut toutefois caractériser l'ensemble de non-différentiabilité de d_K . En suivant [7], on définit :

Définition 2.3.1. L'axe médian d'un compact $K \subset \mathbb{R}^d$ est l'ensemble

$$\operatorname{Med}(K) = \{x \in \mathbb{R}^d \text{ tel que } x \text{ admette au moins deux projections sur } K\}.$$

Par exemple, l'axe médian d'une ellipse et d'un rectangle sont :



On peut définir une fonction continue $\operatorname{proj}_K : \mathbb{R}^d \setminus \operatorname{Med}(K) \rightarrow K$.

Lemme 2.3.2. Soit k une projection sur K d'un point $x \in \mathbb{R}^d$ quelconque. Alors pour tout $h \in \mathbb{R}^d$,

$$d_K^2(x + h) \leq d_K^2(x) + 2\langle x - k, h \rangle + \|h\|^2.$$

En particulier, si d_K^2 est différentiable en x , alors $\nabla d_K^2(x) = 2(x - k)$.

Démonstration. Puisque $d_K(x+h) \leq \|x+h-k\|$, on a :

$$\begin{aligned} d_K(x+h)^2 &\leq \|x-k\|^2 + 2\langle x-k, h \rangle + \|h\|^2 \\ &= d_K^2(x) + 2\langle x-k, h \rangle + \|h\|^2. \end{aligned}$$

Maintenant, si d_K^2 est différentiable en x , on a que $d_K^2(x+h) = d_K^2(x) + \langle \nabla d_K^2(x), h \rangle + o(\|h\|)$. Combinée avec l'inégalité précédente, on obtient :

$$d_K^2(x) + \langle \nabla d_K^2(x), h \rangle + o(\|h\|) \leq d_K^2(x) + 2\langle x-k, h \rangle + \|h\|^2.$$

Cette inégalité n'est vérifiée pour tout $h \in \mathbb{R}^d$ que si $\nabla d_K^2(x) = 2(x-k)$. \square

Proposition 2.3.3. L'ensemble des points de non-différentiabilité de d_K sur $\mathbb{R}^d \setminus K$ est exactement $\text{Med}(K)$.

Démonstration. Si x est un point de différentiabilité de d_K , c'est aussi le cas pour d_K^2 . Le lemme précédent montre alors que l'on a $\nabla d_K^2(x) = 2(x-k)$ pour toute projection k de x sur K . Le point k est donc en fait unique. On a donc montré que $x \notin \text{Med}(K)$.

Nous renvoyons à [7] pour le sens réciproque. \square

2.3.2 Cellules de Voronoï

Soit X un nuage fini de points de \mathbb{R}^d . L'axe médian de X admet une structure particulière. Pour $x \in X$, on définit la cellule de Voronoï de x comme l'ensemble

$$\text{Vor}(x) = \{y \in \mathbb{R}^d, \forall x' \in X, \|y-x\| \leq \|y-x'\|\}.$$

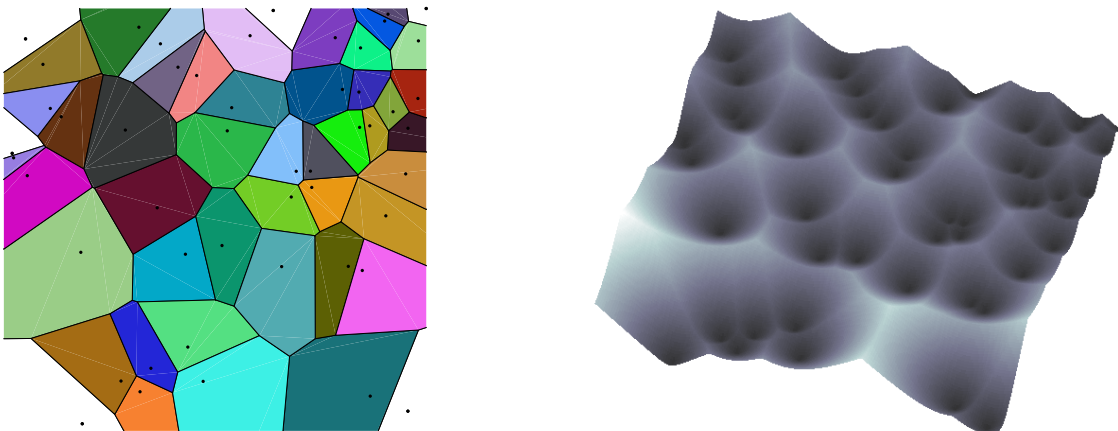
Chacune de ces cellules est un polyèdre, car elles sont l'intersection des demi-espaces définis par les applications affines

$$y \mapsto \|x-y\|^2 - \|x'-y\|^2 = \|x\|^2 - \|x'\|^2 - 2\langle x-x', y \rangle.$$

Leur union vaut l'espace en entier.

Par définition, on voit que les faces des cellules de Voronoï est exactement l'ensemble des points qui admettent au moins deux projections sur X , c'est à dire l'ensemble $\text{Med}(X)$.

Cela s'observe sur les figures suivantes : le diagramme de Voronoï d'un nuage de point de \mathbb{R}^2 , ainsi que la fonction distance associée.



2.3.3 Gradient généralisé

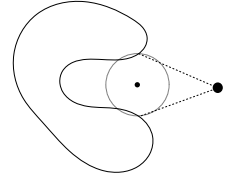
Le gradient ∇d_K n'est défini que sur $\mathbb{R}^d \setminus \text{Med}(K)$. Toutefois, on peut définir un champ de vecteur ∇_K , de norme semi-continue inférieurement, qui coïncide avec ∇d_K là où il est défini.

Soit $x \in \mathbb{R}^d$, et $\Gamma(x)$ l'ensemble de ses projections. On appelle boule englobante de $\Gamma(x)$ une boule de l'espace euclidien qui le contient. Puisque $\Gamma(x)$ est borné, l'ensemble des rayons des boules englobantes est non-vide :

$$\{r \leq 0, \exists y \in \mathbb{R}^d, \Gamma(x) \subset B(y, r)\} \neq \emptyset.$$

Si $B(y_1, r_1)$ et $B(y_2, r_2)$ sont deux telles boules, il existe une autre boule englobante de rayon plus petit que r_1 et r_2 (par exemple une boule dont le diamètre est celui de $B(y_1, r_1) \cap B(y_2, r_2)$). L'ensemble précédent est donc compact.

Ainsi, on peut définir une boule englobante de rayon minimal. On notera $\Theta(x)$ son centre et $R(x)$ son rayon. En particulier, si $x \notin \text{Med}(K)$, on a que $\Gamma(x) = \{\Theta(x)\}$ et $R(x) = 0$.



Définition 2.3.4. Le gradient généralisé de d_K est le champ de vecteurs

$$\nabla_K(x) = \begin{cases} \frac{x - \Theta(x)}{d_K(x)} & \text{si } x \notin K, \\ 0 & \text{sinon.} \end{cases}$$

Proposition 2.3.5. Le gradient généralisé de d_K coïncide avec son gradient là où il est défini.

Démonstration. Si $x \in K$ est un point de différentiabilité de f , alors $\nabla d_K(x) = 0$ car x est un minimum local.

Si $x \in \mathbb{R}^d \setminus K$ est un point de différentiabilité de f , alors $x \notin \text{Med}(K)$, et on a

$$\nabla d_K^2(x) = 2d_K(x)\nabla d_K(x).$$

On déduit du lemme 2.3.2 que $\nabla d_K(x) = \frac{x - \text{proj}_K(x)}{d_K(x)}$. Par définition, on a donc $\nabla_K(x) = \nabla d_K(x)$. On a bien étendu le gradient de d_K . \square

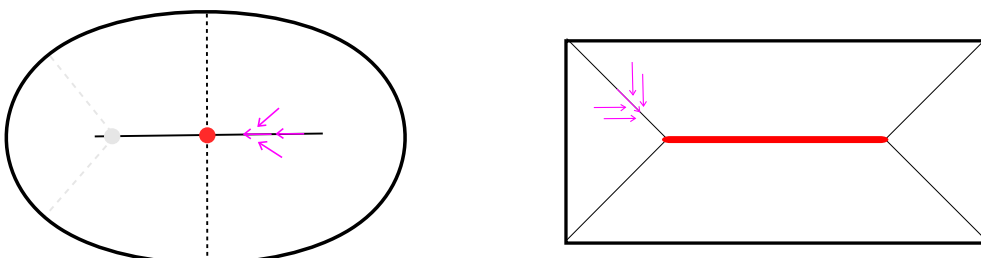
Le résultat suivant est tiré de [8].

Proposition 2.3.6. Soit $x \in \mathbb{R}^d \setminus K$. Alors $\|\nabla_K(x)\|^2 = 1 - \frac{R(x)^2}{d_K(x)^2}$. En particulier, $\forall x \in \mathbb{R}^d$,

$$\begin{cases} \|\nabla_K(x)\| = 0 & \text{si } x \in K, \\ 0 \leq \|\nabla_K(x)\| < 1 & \text{si } x \in \text{Med}(K), \\ \|\nabla_K(x)\| = 1 & \text{sinon.} \end{cases}$$

On appellera points critiques de d_K les points d'annulation du gradient généralisé ∇_K . Par définition, ce sont les points $x \in \mathbb{R}^d$ tels que $\Theta(x) = x$. Autrement dit, ce sont les points qui appartiennent à l'enveloppe convexe de leurs projections. On dira aussi qu'une valeur $c \in \mathbb{R}$ est régulière si $d_K^{-1}(\{c\})$ ne comporte pas de points critiques.

Sur la figure suivante, nous indiquons en rouge les points critiques de la fonction distance à l'ellipse et au carré, ainsi que quelques valeurs du gradient généralisé en rose.



Bien que ∇_K ne soit pas continu, on peut définir un flot $\Phi : \mathbb{R}^+ \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ tel que, pour tout $x \in \mathbb{R}^d$, la courbe $\gamma : t \mapsto \Phi^t(x)$ soit une fonction différentiable à droite et vérifiant $\gamma'(t) = \nabla_K(\gamma(t))$. Tout comme pour les fonctions de Morse, cette déformation permet d'étudier le comportement des sous-niveaux de d_K .

Théorème 2.3.7. Si $c \in \mathbb{R}$ est une valeur régulière, alors $d_K^{-1}(\{c\})$ est une sous-variété \mathcal{C}^0 de \mathbb{R}^d .

Théorème 2.3.8. (Lemme d'isotopie) Soit $c < d \in \mathbb{R}$. Si l'intervalle $[c, d]$ ne contient pas de valeurs critiques, alors les sous-niveaux $K^t, t \in [c, d]$, se rétractent les uns sur les autres.

Autrement dit, le module de persistance \mathbb{V}_{d_K} n'a pour valeurs critiques que les valeurs critiques de d_K . Supposons que les valeurs critiques soient en nombre fini. Alors, comme ce que l'on a dit pour les fonctions de Morse, on a que \mathbb{V}_{d_K} est modéré.

Pour finir, nous définissons :

Définition 2.3.9. Le wfs (pour l'anglais *weak feature size*) du compact K est le réel

$$\text{wfs}(K) = \inf\{c > 0 \text{ tel que } c \text{ soit une valeur critique}\}.$$

Les résultats précédents montrent que les sous-niveaux de d_K conservent le type d'homotopie de K dans l'intervalle $[0, \text{wfs}(K)[$.

2.4 Fonctions presque-distance

Dans cette partie, nous présentons brièvement comment les résultats pour les fonctions distance se transposent à l'identique pour une classe plus générale de fonctions filtrantes.

Définition 2.4.1. On dit que la fonction $f : \mathbb{R}^d \rightarrow \mathbb{R}$ est une presque-distance (de l'anglais *distance-like*) si :

- f est 1-lipschitzienne,
- f est propre,
- et f^2 est 1-semiconcave, c'est à dire que $x \mapsto \|x\|^2 - f^2(x)$ est convexe.

Sans plus de détails, nous affirmons que la semiconcavité de f^2 permet de définir un gradient généralisé de f . Comme pour les fonctions distance, on peut ensuite montrer l'existence d'un flot contre ce champ de vecteurs. On obtient finalement :

Théorème 2.4.2. (Lemme d'isotopie) Soit f une fonction presque-distance et $c < d \in \mathbb{R}$. Si l'intervalle $[c, d]$ ne contient pas de valeurs critiques de son gradient généralisé, alors les ensembles de niveaux $f^t, t \in [c, d]$ sont isotopes.

2.5 Stabilité des modules

2.5.1 Régularité

Proposition 2.5.1. Soit $K \subset \mathbb{R}^d$ un compact et $d_K : \mathbb{R}^d \rightarrow \mathbb{R}$ la fonction distance à K . Alors le module de persistance \mathbb{V}_{d_K} est q -modéré.

Démonstration. Soit $\epsilon > 0$ et $t \in \mathbb{R}$. Puisque le sous-niveau K^t est compact, définissons $P \subset K^t$ un sous-ensemble fini ϵ -dense. On a les inclusions :

$$K^t \subset P^\epsilon \subset K^{t+\epsilon},$$

qui se traduisent au niveau homologique par le diagramme commutatif :

$$\begin{array}{ccccc} H_*(K^t) & \longrightarrow & H_*(P^\epsilon) & \longrightarrow & H_*(K^{t+\epsilon}) \\ & & \searrow v_t^{t+\epsilon} & \nearrow & \end{array}$$

On verra (théorème 3.3.3) que l'espace vectoriel $H_*(P^{t+\epsilon})$ est de dimension finie. Par conséquent, l'application $v_t^{t+\epsilon}$ est de rang fini.

Ceci étant valable pour tout $t \in \mathbb{R}$ et $\epsilon > 0$, on conclut que le module en question est q -modéré. \square

En fait ce résultat se généralise directement :

Proposition 2.5.2. Soit $f : \mathbb{R}^d \rightarrow \mathbb{R}$ une fonction propre et minorée. Alors le module de persistance \mathbb{V}_f est q -modéré.

Démonstration. Soit $s < t \in \mathbb{R}$. Puisque f est propre, les ensembles de niveaux $f^{-1}(\{s\})$ et $f^{-1}(\{t\})$ et compacts. De plus, par continuité de f , la valeur

$$\delta = \inf \{ \|x - y\|, x \in f^{-1}(\{s\}), y \in f^{-1}(\{t\}) \}$$

est strictement positive.

Soit $T \subset \mathbb{R}^d$ un d -simplexe régulier qui contient f^t . Nous pouvons effectuer des subdivisions barycentriques successives de T , jusqu'à obtenir une triangulation (finie) de T par des simplexes $\{T_i\}_{i \in I}$ de diamètres strictement plus petits que δ .

Par définition de δ , ces simplexes vérifient la propriété de ne pas intersecter à la fois $f^{-1}(\{s\})$ et $f^{-1}(\{t\})$.

Soit alors $\{T_j\}_{j \in J}$ l'ensemble des simplexes qui intersectent f^s , et $K = \bigcup_{j \in J} T_j$ leur union. On a les inclusions :

$$f^s \subset K \subset f^t,$$

qui se transportent par fonctorialité :

$$\begin{array}{ccccc} H_*(f^s) & \longrightarrow & H_*(K) & \longrightarrow & H_*(f^t) \\ & & \searrow v_s^t & \nearrow & \end{array}$$

Or l'espace vectoriel $H_*(K)$ est de dimension finie car c'est un espace triangulé (son homologie singulière coïncide donc avec son homologie simpliciale en tant que complexe simplicial). L'application v_s^t est par conséquent de rang fini. \square

2.5.2 Stabilité

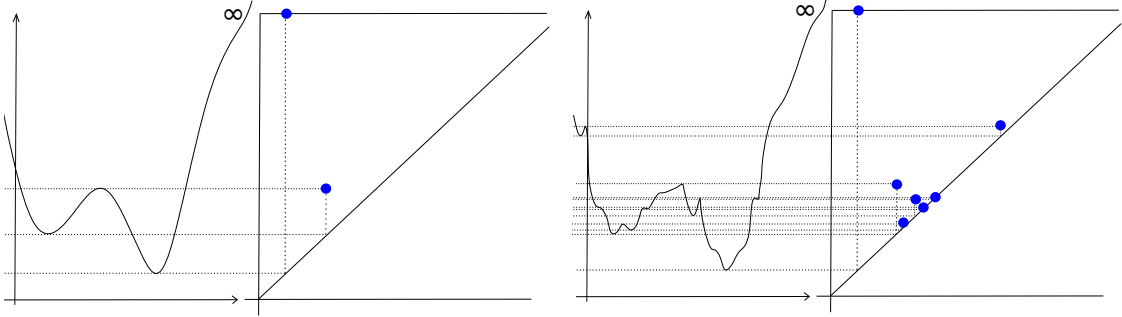
Proposition 2.5.3. Soient $f, g : E \rightarrow \mathbb{R}$ deux fonctions continues définies sur un espace topologique, et $\epsilon = \|f - g\|_\infty$. Alors les modules de persistance \mathbb{V}_f et \mathbb{V}_g sont ϵ -entrelacés.

Démonstration. Il suffit de remarquer que, pour tout $t \in \mathbb{R}$, on a les inclusions $f^t \subset g^{t+\epsilon} \subset f^{t+2\epsilon}$. On obtient bien les diagrammes voulus entre les groupes d'homologie. \square

Si les diagrammes de persistance de ces modules sont bien définis, alors le théorème d'isométrie permet de majorer leur distance bottleneck :

$$d_b(\mathbb{V}_f, \mathbb{V}_g) \leq \|f - g\|_\infty.$$

Pour illustrer cette propriété, nous représentons le diagramme de persistance des sous-niveaux d'une fonction de \mathbb{R}^2 dans \mathbb{R} , ainsi que son diagramme lorsqu'elle est soumise à de petites perturbations en distance de Hausdorff. On voit apparaître des points proches de la diagonale.



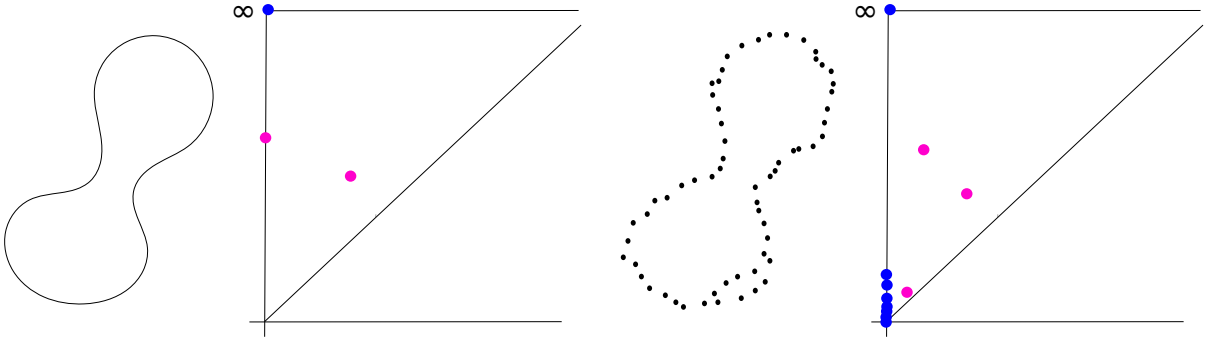
Soient maintenant $K, L \subset \mathbb{R}^d$ deux compacts. Leur distance de Hausdorff peut être définie comme la valeur

$$d_H(K, L) = \|d_L - d_K\|_\infty.$$

Par la proposition 2.5.1, on sait que les digrammes de persistance des modules \mathbb{V}_{d_K} et \mathbb{V}_{d_L} sont bien définis. On a donc :

$$d_b(\mathbb{V}_{d_K}, \mathbb{V}_{d_L}) \leq d_H(K, L).$$

Ce résultat permet de justifier la proximité des diagrammes d'une sous-variété et d'un nuage de points qui l'échantillonne :



3 Complexes simpliciaux filtrés

3.1 Filtrations

Filtration d'un complexe simplicial. Soit k un complexe simplicial abstrait (éventuellement infini), et V son ensemble de sommets. On appelle filtration de K , et on note \mathbb{K} , une suite de sous-complexes

$$\emptyset = K^0 \subset K^1 \subset \dots \subset K.$$

L'homologie simpliciale de ces complexes donne lieu à une suite croissante d'espaces vectoriels $V_i = H_*(K^i)$. Les inclusions entre les sous-complexes simpliciaux induisent des morphismes $v_i^j : V_i \rightarrow V_j$ qui forment une structure de module de persistance. On notera dans la suite $\mathbb{V} = ((V_i)_{i \in \mathbb{N}}, (v_i^j)_{i \leq j \in \mathbb{N}})$ ce module.

Si la famille est finie, on note :

$$\emptyset = K^0 \subset K^1 \subset \dots \subset K^n = K.$$

Nombres de Betti persistants. Considérons le module de persistance issu des k -èmes groupes d'homologie du complexe simplicial filtré fini \mathbb{K} . Dans ce contexte, les rangs de $v_i^j : H_k(K^i) \rightarrow H_k(K^j)$ sont appelés les nombres de Betti persistants, notés $\beta_k^{i,j}$. Rappelons que dans ce cas, ces nombres suffisent à caractériser le module de persistance à isomorphisme près (proposition 1.4.10).

Le nombre $\beta_k^{i,j}$ représente le nombre de classes d'homologie de K^i qui ne sont toujours pas des bords dans K^j . Plus précisément, on écrit

$$v_i^j : Z_k(K^i) / B_k(K^i) \rightarrow Z_k(K^j) / B_k(K^j).$$

Le noyau de cette application est formé par les cycles de $Z_k(K^i)$ qui sont des bords dans K^j . Autrement dit, $\ker v_i^j = Z_k(K^i) \cap B_k(K^j)$. Par le théorème d'isomorphisme, on obtient :

$$\text{Im}(v_i^j) \simeq Z_k(K^i) / Z_k(K^i) \cap B_k(K^j),$$

et le nombre de Betti persistant se réécrit alors

$$\beta_k^{i,j} = \dim Z_k(K^i) / Z_k(K^i) \cap B_k(K^j).$$

3.2 Algorithme de persistance

Les résultats de cette partie proviennent de [9].

3.2.1 Calcul des nombres de Betti

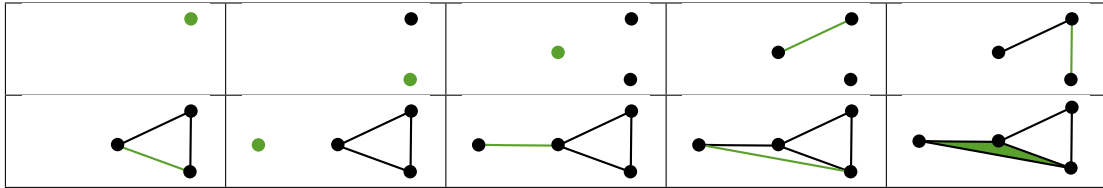
On considère toujours le complexe simplicial filtré fini

$$\emptyset = K^0 \subset K^1 \subset \dots \subset K^n = K.$$

On suppose de plus que les simplexes sont ajoutés un par un : pour tout $i \in \llbracket 1, n \rrbracket$,

$$K^i = K^{i-1} \cup \{\sigma^i\}.$$

La figure suivante est un exemple d'une telle filtration :



On dira qu'un simplexe $\sigma \in K^i$ est contenu dans une chaîne $c = \sum_{\nu \in K^i} \epsilon_\nu \nu$ de K^i si $\epsilon_\sigma = 1$.

On dira aussi que deux chaînes c, c' de K^i sont homologues dans K^i si leur différence est un bord de K^i .

Soit d la dimension de K . Nous allons décrire un algorithme qui permet de calculer les nombres de Betti β_0, \dots, β_d de K .

Définition 3.2.1. Soit $\sigma^i \in K^i \setminus K^{i-1}$ et k sa dimension. On dira que σ^i est :

- positif s'il est contenu dans un k -cycle de K^i ,
- négatif sinon.

Le lemme suivant montre qu'un simplexe positif crée une nouvelle classe de $H_k(K^i)$ (linéairement indépendante l'image de $H_k(K^{i-1}) \hookrightarrow H_k(K^i)$).

Lemme 3.2.2. Supposons que σ^i soit positif, contenu dans le k -cycle c de K^i . Alors c n'est pas homologue (dans K^i) à un k -cycle de K^{i-1} .

Démonstration. Puisque $(K_j)_j$ est une filtration, on sait déjà que σ^i n'est pas la face d'un simplexe de K^i (on vient de l'ajouter). En particulier, σ^i n'est contenu dans aucun bord de K^i . Maintenant, supposons par l'absurde qu'il existe un k -cycle c' de K^{i-1} qui soit homologue à c . Alors $c + c'$ est un bord. Mais puisque σ^i est contenu dans $c + c'$, on est en contradiction avec la remarque précédente. \square

Le lemme suivant montre qu'un simplexe négatif détruit une classe de $H_k(K^{i-1})$.

Lemme 3.2.3. Supposons que σ^i soit négatif. Alors le cycle $\partial\sigma^i$ n'est pas un bord de K^{i-1} .

Démonstration. Puisque $(K_j)_j$ est une filtration, $\partial\sigma^i$ est bien inclus dans K^{i-1} . Supposons par l'absurde que $\partial\sigma^i$ soit un bord de K^{i-1} , c'est à dire qu'il existe une k -chaîne c de K^{i-1} telle que $\partial\sigma^i = \partial c$. On obtient $\partial(c + \sigma^i) = 0$, et donc un cycle $c + \sigma^i$ qui contient σ^i . Cela contredit la négativité de σ^i . \square

On est maintenant capable de suivre l'évolution de l'homologie de K au cours de la filtration :

Proposition 3.2.4. On considère le simplexe $\sigma^i \in K^i \setminus K^{i-1}$ et k sa dimension.

- si σ^i est positif, alors $\beta_k(K^i) = \beta_k(K^{i-1}) + 1$,

— sinon, $\beta_{k-1}(K^i) = \beta_{k-1}(K^{i-1}) - 1$.

Démonstration. Si σ^i est positif : nous venons de voir que dans ce cas, $\beta_k(K^i) \geq \beta_k(K^{i-1}) + 1$. Pour montrer l'égalité, choisissons deux k -cycles c et c' de K^i qui contiennent σ^i . Le cycle $c + c'$ ne contient pas σ^i , c'est donc un cycle de K^{i-1} . La famille (c, c') n'est donc pas linéairement indépendante à une base de $H_k(K^{i-1})$. On a donc $\dim H_k(K^i) \leq \dim H_k(K^{i-1}) + 1$.
Si σ^i est négatif : nous avons aussi vu que $\beta_{k-1}(K^i) \leq \beta_{k-1}(K^{i-1}) - 1$. Pour montrer l'égalité, il suffit de remarquer que deux cycles homologues à $\partial\sigma^i$ dans K^i sont aussi homologues entre eux. On a donc $\dim H_k(K^i) \geq \dim H_k(K^{i-1}) - 1$, ce qui conclut. \square

Ainsi, si l'on est en mesure de tester la positivité ou négativité d'un simplexe (voir remarque 3.2.9), alors l'algorithme suivant permet de calculer les nombres de Betti désirés.

Data: Un complexe simplicial filtré K de dimension d composé de n simplexes.

Result: Ses nombres de Betti β_0, \dots, β_d .

$\beta_0 \leftarrow 0, \dots, \beta_d \leftarrow 0$;

for $i \leftarrow 1$ **to** n **do**

$k = \dim(\sigma^i)$;

if σ^i est positif **then**

$\beta_k(K^i) \leftarrow \beta_k(K^i) + 1$;

else

$\beta_{k-1}(K^i) \leftarrow \beta_{k-1}(K^{i-1}) - 1$;

end

end

3.2.2 Nombres de Betti persistants

Pour calculer les nombres de Betti persistants de la filtration \mathbb{K} , nous allons adapter l'algorithme précédent, pour qu'il prenne en compte les temps de naissance et de mort des cycles. Pour cela, nous allons construire, pour tout $i \leq n$ et $k \leq d$, un ensemble de cycles

$$\mathcal{B}_k^i = \{c_1, \dots, c_{n_i}\},$$

qui forment une base de $H_k(K^i)$, et tels que tout changement de base \mathcal{B}_k^i à \mathcal{B}_k^{i+1} se fasse par ajout ou suppression d'un cycle (en fonction de la positivité ou négativité du simplexe ajouté).

Le lemme suivant, de [9], donne un choix judicieux pour ces cycles :

Lemme 3.2.5. Soit $\sigma^i \in K^i \setminus K^{i-1}$ un k simplexe positif. Alors il existe un unique k -cycle c de K^i tel que :

- c ne soit pas un bord dans K^i ,
- c contienne σ^i , et
- c ne contienne pas d'autre simplexe positif.

Pour un simplexe positif σ^i , on notera alors c_i cet unique cycle associé.

Décrivons maintenant l'algorithme. Soit $i \leq 0$, et supposons que les bases \mathcal{B}_l^{i-1} de $H_l(K^{i-1})$ aient été construites ($l = 1, \dots, d$). On a affaire au nouveau simplexe $\sigma^i \in K^i \setminus K^{i-1}$. Soit k sa dimension.

Si σ^i est positif, alors il correspond à un unique cycle c défini par la proposition 3.2.5, que l'on ajoute à la base \mathcal{B}_k^{i-1} :

$$\mathcal{B}_k^i = \mathcal{B}_k^{i-1} \cup \{c_i\}$$

Les résultats de la partie 3.2.1 montrent que \mathcal{B}_k^i est bien une base de $H_k(K^i)$.

Si σ^i est négatif, nous allons retirer un simplexe de la base \mathcal{B}_{k-1}^{i-1} . Notons $\mathcal{B}_{k-1}^{i-1} = \{\nu_{i_1}, \dots, \nu_{i_J}\}$. D'après le lemme 3.2.3, $\partial\sigma^i$ n'est pas un bord de K^{i-1} . On peut donc l'écrire :

$$\partial\sigma^i = \sum_{j=i_1, \dots, i_J} \epsilon_j \nu_j + b,$$

où les ϵ_j sont uniques et non tous nuls, et b est un bord de K^{i-1} . Soit $l(i)$ le plus grand indice tel que $\epsilon_{l(i)}$ soit non nul. Alors $\sigma^{l(i)}$ est un simplexe positif, et on choisit de retirer son cycle correspondant de \mathcal{B}_{k-1}^{i-1} :

$$\mathcal{B}_{k-1}^i = \mathcal{B}_{k-1}^{i-1} \setminus \{c_{l(i)}\}.$$

De même, \mathcal{B}_{k-1}^i est bien une base de $H_{k-1}(K^i)$

Remarque 3.2.6. L'indice $l(i)$ ainsi défini répond au principe suivant : si un simplexe doit effacer un cycle qui s'écrit comme somme de plusieurs éléments de la base \mathcal{B}_{k-1}^{i-1} , alors on choisit de supprimer l'élément qui est arrivé le plus tard.

Voici l'algorithme écrit en pseudo-code :

Data: Un complexe simplicial filtré K de dimension d composé de n simplexes.

Result: Les ensembles de paires persistantes L_1, \dots, L_d .

$L_1 \leftarrow \emptyset, \dots, L_d \leftarrow \emptyset;$

for $i \leftarrow 1$ **to** n **do**

$k = \dim(\sigma^i);$

if σ^i est négatif **then**

$L_k \leftarrow L_k \cup \{(\sigma^{l(i)}, \sigma^j)\};$

end

end

Définition 3.2.7. Les couples $(\sigma^{l(j)}, \sigma^j)$ ainsi définis sont appelés paires persistantes.

Réciproquement, pour tout indice $i \in \llbracket 1, n \rrbracket$ tel que σ^i soit un simplexe positif, on note $c(i)$ l'indice du simplexe $\sigma^{c(i)}$ tel que $(\sigma^i, \sigma^{c(i)})$ soit une paire persistante. Si un tel indice n'existe pas, on pose $c(i) = +\infty$.

Par construction de l'algorithme, on obtient donc les nombres de Betti persistants par

$$\beta_k^{i,j} = \{i \leq a \text{ tel que } \sigma^i \text{ soit un } k\text{-simplexe positif et } c(i) > b\}.$$

On en déduit :

Proposition 3.2.8. Le module de persistance \mathbb{V} du k -ème groupe d'homologie se décompose en

$$\mathbb{V} = \bigoplus \mathbb{I}([i, c(i)]),$$

où la somme porte sur les indices i tels que σ^i soit un k -simplexe positif.

Démonstration. On sait que le module \mathbb{V} se décompose en une somme finie $\mathbb{V} = \bigoplus \mathbb{I}(a_i, b_i)$ sur \mathbb{N} . D'après la proposition 1.4.10, le nombre d'intervalles $([a, b])$ dans cette décomposition est :

$$\beta_k^{a,b-1} - \beta_k^{a,b} + \beta_k^{a-1,b} - \beta_k^{a-1,b-1}.$$

Or

$$\beta_k^{a,b-1} - \beta_k^{a,b} = \{i \leq a \text{ tel que } \sigma^i \text{ soit un } k\text{-simplexe positif et } c(i) = b\},$$





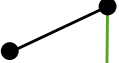
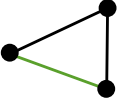


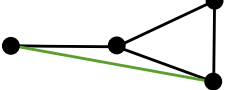

$$\beta_k^{a-1,b} - \beta_k^{a-1,b-1} = \{i \leq a-1 \text{ tel que } \sigma^i \text{ soit un } k\text{-simplexe positif et } c(i) = b\}.$$

On déduit $\beta_k^{a,b-1} - \beta_k^{a,b} + \beta_k^{a-1,b} - \beta_k^{a-1,b-1} = \begin{cases} 1 & \text{si } \sigma^a \text{ est un } k\text{-simplexe positif tel que } c(i) = b, \\ 0 & \text{sinon.} \end{cases}$

□

Remarque 3.2.9. Tel quel, l’algorithme n’est pas utilisable puisqu’il faut encore savoir tester la positivité d’un simplexe et trouver l’indice $l(i)$ en question. En fait, une des versions actuellement utilisées de cet algorithme consiste à transformer le problème en la réduction d’une matrice de bord.

Nous présentons maintenant une réalisation de cette algorithme. Le dernier simplexe ajouté est représenté en vert.

i	K^i	Positivité	β_0	\mathcal{B}_0^i	β_1	\mathcal{B}_1^i
1		+	1	σ_1	0	
2		+	2	σ_1, σ_2	0	
3		+	3	$\sigma_1, \sigma_2, \sigma_3$	0	
4		−	2	σ_1, σ_3	0	
5		−	1	σ_1	0	
6		+	1	σ_1	1	$\sigma_4 + \sigma_5 + \sigma_6$
7		+	2	σ_1, σ_7	1	$\sigma_4 + \sigma_5 + \sigma_6$
8		−	1	σ_1	1	$\sigma_4 + \sigma_5 + \sigma_6$
9		+	1	σ_1	2	$\sigma_4 + \sigma_5 + \sigma_6,$ $\sigma_4 + \sigma_5 + \sigma_8 + \sigma_9$
10		−	1	σ_1	1	$\sigma_4 + \sigma_5 + \sigma_6$

3.3 Complexe de Čech

3.3.1 Théorème du nerf persistant

Souvent, la construction d'un complexe simplicial est un intermédiaire permettant de calculer l'homologie persistante d'un espace topologique sous-jacent. Dans cette partie, nous discuterons du théorème du nerf, qui permet d'établir un pont entre ces deux espaces.

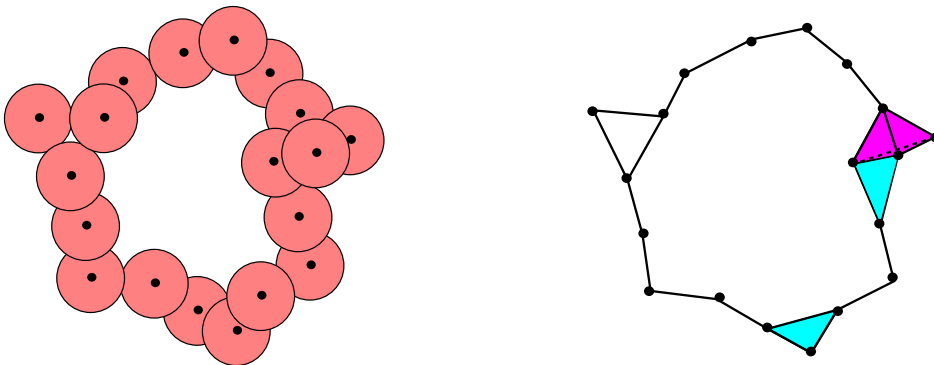
Considérons un espace topologique E , et $\mathcal{U} = \{U_i\}_{i \in I}$ un recouvrement ouvert (éventuellement infini).

Définition 3.3.1. On dit que le recouvrement \mathcal{U} est un bon recouvrement si chaque intersection finie $\bigcap_{j \in J \subset I} U_j$ est soit vide soit contractile.

En particulier, si E est un espace métrique et les U_i des convexes, alors \mathcal{U} est un bon recouvrement.

Définition 3.3.2. Le nerf du recouvrement \mathcal{U} , noté $\mathcal{N}(\mathcal{U})$, est un complexe simplicial abstrait, construit sur l'ensemble de sommets I , et tel qu'un sous-ensemble $J \subset I$ soit un simplexe de $\mathcal{N}(\mathcal{U})$ si et seulement $\bigcap_{j \in J \subset I} U_j \neq \emptyset$

La figure suivante représente le nerf d'un recouvrement par des boules. On voit apparaître dans le complexe simplicial six simplexes de dimension 2 et un de dimension 3.



Énonçons tout de suite le théorème du nerf dans sa version classique (par exemple dans [10]) :

Théorème 3.3.3. Si E est paracompact et \mathcal{U} est un bon recouvrement, alors $\mathcal{N}(\mathcal{U})$ est du même type d'homotopie que E .

Autrement dit, si l'on a calculé les groupes d'homologie simpliciale du nerf d'un de ses bon recouvrements \mathcal{U} , alors on connaît aussi les groupes d'homologie singulière de l'espace E .

L'application classique de ce théorème est la suivante : soit X^t le voisinage tubulaire d'un ensemble fini de points $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$. Alors X^t est recouvert par $\mathcal{U} = \{B(x_i, t)\}_{i \leq n}$. Dans ce cadre, le nerf de \mathcal{U} est appelé complexe de Čech de X et paramètre t . On le note

$$\check{\text{Cech}}^t(X) := \mathcal{N}(\mathcal{U}).$$

Le théorème du nerf fournit une équivalence d'homotopie, c'est à dire deux applications continues $X^t \xrightarrow{\phi} \check{\text{Cech}}^t(X)$ et $\check{\text{Cech}}^t(X) \xrightarrow{\psi} X^t$, telles que l'on obtienne les diagrammes suivants au niveau homologique :

$$\begin{array}{ccc} H_*(X^t) & \xrightarrow{\Phi} & H_*(\check{\text{Cech}}^t(X)) \\ & \searrow \text{id} & \downarrow \Psi \\ & & H_*(X^t) \end{array} \quad \text{et} \quad \begin{array}{ccc} H_*(\check{\text{Cech}}^t(X)) & \xrightarrow{\Psi} & H_*(X^t) \\ & \searrow \text{id} & \downarrow \Phi \\ & & H_*(\check{\text{Cech}}^t(X)) \end{array}$$

Maintenant, nous voulons comparer les structures de modules de persistance $(H_*(X^t))_{t \geq 0}$ et $H_*(\check{\text{Cech}}^t(X))_{t \geq 0}$. Faisons varier le paramètre t . Pour tout $s \in \mathbb{R}^+$, le théorème donne aussi des applications ϕ_s et ψ_s comme précédemment. Pour montrer un isomorphisme entre ces deux modules, il suffirait d'avoir des diagrammes commutatifs

$$\begin{array}{ccc} X^s & \xrightarrow{\quad} & X^t \\ \downarrow \phi_s & & \downarrow \phi_t \\ \check{\text{Cech}}^s(X) & \xrightarrow{\quad} & \check{\text{Cech}}^t(X) \end{array} \quad \text{et} \quad \begin{array}{ccc} X^s & \xrightarrow{\quad} & X^t \\ \psi_s \uparrow & & \psi_t \uparrow \\ \check{\text{Cech}}^s(X) & \xrightarrow{\quad} & \check{\text{Cech}}^t(X) \end{array}$$

ce qui n'est pas donné a priori.

Le lemme suivant, de [11], indique comment choisir de bonnes homotopies :

Lemme 3.3.4. Soient $E \subset E'$ deux espaces paracompacts, et $\mathcal{U} = \{U_i\}_{i \in I}$, $\mathcal{U}' = \{U'_i\}_{i \in I}$ deux bon recouvrements ouverts tels que $\forall i \in I, U_i \subset U'_i$. Alors il existe deux équivalences d'homotopie $E \xrightleftharpoons[\psi]{\phi} \mathcal{N}(\mathcal{U})$ et $E' \xrightleftharpoons[\psi']{\phi'} \mathcal{N}(\mathcal{U}')$ telles que les diagrammes suivants commutent :

$$\begin{array}{ccc} E & \xrightarrow{\quad} & E' \\ \downarrow \phi & & \downarrow \phi' \\ \mathcal{N}(\mathcal{U}) & \xrightarrow{\quad} & \mathcal{N}(\mathcal{U}') \end{array} \quad \text{et} \quad \begin{array}{ccc} E & \xrightarrow{\quad} & E' \\ \psi \uparrow & & \psi' \uparrow \\ \mathcal{N}(\mathcal{U}) & \xrightarrow{\quad} & \mathcal{N}(\mathcal{U}') \end{array}$$

En fait, la démonstration de ce lemme donne même un résultat plus précis :

Théorème 3.3.5. Les modules de persistance $(H_*(X^t))_{t \geq 0}$ et $H_*(\check{\text{Cech}}^t(X))_{t \geq 0}$ sont isomorphes.

3.3.2 Complexe de Vietoris-Rips

Soit $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$. On vient de voir que le calcul de la persistance de ce nuage de points se réduit à celui de son complexe de Čech. Toutefois, en pratique, la construction de ce complexe requiert le prédicat géométrique :

Est-ce que les n boules $B(x_1, t), \dots, B(x_n, t)$ s'intersectent ?

Ceci n'est pas aisément implémentable, et on préfère parfois une version approchée du complexe de Čech : le complexe de Vietoris-Rips, $\text{Rips}^t(X)$.

Définition 3.3.6. Le complexe de Rips $\text{Rips}^t(X)$ est le complexe simplicial d'ensemble de sommets X et tel que pour tout $I \subset X$, $I \in \text{Rips}^t(X)$ si et seulement si pour tout $\{a, b\} \subset I$, $\|b - a\| \leq t$.

En fait, $\text{Rips}^t(X)$ est le complexe de clique du 1-squelette du complexe de Čech $\check{\text{Cech}}^{\frac{t}{2}}(X)$:

Définition 3.3.7. Soit G un graphe d'ensemble de sommets V . Le complexe de clique de G est le complexe simplicial $X(G)$ construit sur V tel que pour tout sous-ensemble $I \subset V$, $I \in X(G)$ si et seulement si toutes les paires $\{a, b\} \subset I$ sont des arêtes de G .

On peut observer à gauche un complexe simplicial, et à droite le complexe de Rips correspondant :



On peut comparer ces deux complexes de la façon suivante :

Proposition 3.3.8. Pour tout $t \geq 0$, on a les inclusions suivantes entre complexes simpliciaux :

$$\check{\text{Cech}}^{\frac{t}{2}}(X) \subset \text{Rips}^t(X) \subset \check{\text{Cech}}^t(X)$$

Démonstration. La première inclusion vient de la remarque précédente : si un simplexe $I \subset X$ appartient à $\check{\text{Cech}}^{\frac{t}{2}}(X)$, alors en particulier ses points sont à distance plus petite que $\frac{t}{2} + \frac{t}{2} = t$. I est donc aussi un simplexe de $\text{Rips}^t(X)$.

Pour la seconde, soit I un simplexe de $\text{Rips}^t(X)$, et $x \in I$ un point quelconque. Par définition, tout autre point $y \in I$ vérifie $\|x - y\| \leq t$. Autrement dit, chaque boule $B(y, t)$ contient x . I est donc aussi un simplexe de $\check{\text{Cech}}^t(X)$. \square

On obtient ainsi des morphismes entre les groupes d'homologie :

$$\begin{array}{ccc} H_*(\check{\text{Cech}}^t(X)) & \longrightarrow & H_*(\check{\text{Cech}}^{2t}(X)) \\ \downarrow & \nearrow & \\ H_*(\text{Rips}^{2t}(X)) & & \end{array} \quad \text{et} \quad \begin{array}{ccc} H_*(\text{Rips}^t(X)) & \longrightarrow & H_*(\text{Rips}^{2t}(X)) \\ \downarrow & \nearrow & \\ H_*(\check{\text{Cech}}^t(X)) & & \end{array}$$

Selon notre définition, ceci n'est pas un entrelacement entre les modules $H_*(\check{\text{Cech}}^t(X))_{t \geq 0}$ et $H_*(\text{Rips}^t(X))_{t \geq 0}$. On pourrait toutefois considérer que c'est un entrelacement en échelle logarithmique :

Définition 3.3.9. Soit $\mathbb{U} = (U_t, u_s^t)_{s \leq t \in \mathbb{R}^+}$ et $\mathbb{V} = (V_t, v_s^t)_{s \leq t \in \mathbb{R}^+}$ deux modules de persistance. On appelle ϵ -entrelacement en échelle logarithmique entre \mathbb{U} et \mathbb{V} tout ϵ -entrelacement entre les modules de persistance $\mathbb{U}^{\log} = (U_{e^t})_{t \in \mathbb{R}}$ et $\mathbb{V}^{\log} = (V_{e^t})_{t \in \mathbb{R}}$.

La distance d'entrelacement en échelle logarithmique est notée $d_i^{\log}(\mathbb{U}, \mathbb{V})$. On peut vérifier qu'elle correspond à la distance bottleneck entre les images des diagrammes de persistance de \mathbb{U} et \mathbb{V} par l'application $(a, b) \mapsto (\log(a), \log(b))$.

Dans notre cas, les diagrammes commutatifs précédents se transforment, par le changement de variable $t \mapsto e^t$, en :

$$\begin{array}{ccc} H_*(\check{\text{Cech}}(X))_t^{\log} & \longrightarrow & H_*(\check{\text{Cech}}(X))_{\log(2)+t}^{\log} \\ \downarrow & \nearrow & \\ H_*(\text{Rips}(X))_{\log(2)+t}^{\log} & & \end{array} \quad \text{et} \quad \begin{array}{ccc} H_*(\text{Rips}(X))_t^{\log} & \longrightarrow & H_*(\text{Rips}(X))_{\log(2)+t}^{\log} \\ \downarrow & \nearrow & \\ H_*(\check{\text{Cech}}(X))_t^{\log} & & \end{array}$$

Autrement dit, $H_*(\check{\text{Cech}}(X))_t^{\log}$ et $H_*(\text{Rips}(X))_t^{\log}$ sont $\log(2)$ -entrelacés. Par définition, on dit donc que $H_*(\check{\text{Cech}}^t(X))$ et $H_*(\text{Rips}^t(X))$ sont $\log(2)$ -entrelacés en échelle logarithmique.

3.4 Stabilité

Nous allons développer dans cette partie quelques outils permettant de comparer les complexes de Čech et Rips lorsqu'on les perturbe.

3.4.1 Applications ϵ -simpliciales

Les résultats suivants sont tirés de [12]. On considère deux complexes simpliciaux K et L , X et Y leurs ensemble de sommets.

Définition 3.4.1. Une application simpliciale entre K et L est une application $f : X \rightarrow Y$ telle que l'image de tout simplexe de K soit un simplexe de L .

Soient maintenant $\mathbb{K} = (K^t)_{t \in \mathbb{R}}$ et $\mathbb{L} = (L^t)_{t \in \mathbb{R}}$ des filtrations de K et L .

Définition 3.4.2. Pour $\epsilon \geq 0$, une application ϵ -simpliciale entre \mathbb{K} et \mathbb{L} est une application $f : X \rightarrow Y$ telle pour tout $t \in \mathbb{R}$, l'application induite $f : K^t \rightarrow L^{t+\epsilon}$ soit simpliciale.

Notons dès à présent qu'une telle application induit un ϵ -morphisme entre les modules de persistance $H_*(\mathbb{K})$ et $H_*(\mathbb{L})$.

Dans suite nous allons travailler avec des applications multivaluées. Une application multivaluée entre X et Y , notée $C : X \rightrightarrows Y$, est un sous-ensemble $C \subset X \times Y$ telle que $\pi_X(C) = X$, où $\pi_X : X \times Y \rightarrow X$ est la projection. Si $\sigma \subset X$, son image par C est l'ensemble

$$C(\sigma) = \{y \text{ tel que } (x, y) \in C \text{ et } x \in \sigma\}.$$

On dira qu'une application (univoque) $f : X \rightarrow Y$ est subordonnée à C si pour tout $x \in X$, $(x, f(x)) \in C$. Dans ce cas, on note $f : X \xrightarrow{C} Y$.

Définition 3.4.3. Une application multivaluée ϵ -simpliciale entre \mathbb{K} et \mathbb{L} est une application multivaluée $C : X \rightrightarrows Y$ telle que, pour tout $t \in \mathbb{R}$ et tout simplexe $\sigma \in K^t$, tout sous-ensemble fini de $C(\sigma)$ soit un simplexe de $L^{t+\epsilon}$.

Proposition 3.4.4. A toute application multivaluée ϵ -simpliciale $C : X \rightrightarrows Y$ correspond un ϵ -morphisme canonique entre les modules de persistance, noté $H(C) : \text{pour tout } t \in \mathbb{R}$,

$$H(C)_t : H_*(\mathbb{K}^t) \rightarrow H_*(\mathbb{L}^{t+\epsilon}).$$

Celui-ci est induit par toute application subordonnée $f : X \xrightarrow{C} Y$.

Démonstration. Soient $f_1, f_2 : X \xrightarrow{C} Y$. Il suffit de montrer qu'elles induisent les mêmes morphismes $H_*(K^t) \rightarrow H_*(\mathbb{L}^{t+\epsilon})$ entre les groupes d'homologie. Il suffit pour cela de montrer que f_1 et f_2 sont contiguës, c'est à dire que pour tout simplexe $\sigma \in K^t$, $f_1(\sigma) \cup f_2(\sigma)$ est un simplexe de $L^{t+\epsilon}$ (proposition A.3.3). Or ceci est donné par la définition précédente. \square

Étant données deux application multivaluées $C : X \rightrightarrows Y$ et $D : Y \rightrightarrows Z$, on définit la composition $D \circ C$ par

$$D \circ C = \{(x, z) \in X \times Z, \exists y \in Y, (x, y) \in C \text{ et } (y, z) \in D\}.$$

Proposition 3.4.5. Soient $\mathbb{K}, \mathbb{L}, \mathbb{M}$ trois complexes simpliciaux filtrés et X, Y, Z leurs ensembles de sommets. Soient aussi $C : X \rightrightarrows Y$ ϵ -simpliciale et $D : Y \rightrightarrows Z$ δ -simpliciale. Alors la composée $D \circ C : X \rightrightarrows Z$ est une application multivaluée $(\epsilon + \delta)$ -simpliciale, et on a :

$$H(D \circ C) = H(D) \circ H(C).$$

Démonstration. Par définition, $D \circ C$ est bien $(\epsilon + \delta)$ -simpliciale. Pour montrer le second points, soient $f : X \xrightarrow{C} Y$ et $g : Y \xrightarrow{D} Z$ des applications subordonnées. Alors l'application composée $g \circ f$ est subordonnée à $D \circ C$. Par functorialité de l'homologie simpliciale, on obtient $H(D \circ C) = H(D) \circ H(C)$. \square

3.4.2 Correspondances

Soit toujours $C : X \rightrightarrows Y$ une application multivaluée. On appellera transposée de C , notée C^T , l'ensemble

$$C^T = \{(y, x) \text{ tel que } (x, y) \in C\}.$$

C^T n'est pas a priori une application multivaluée.

Définition 3.4.6. Si la transposée C^T est bien une application multivaluée, alors on dit que C est une correspondance.

Proposition 3.4.7. Soit C est une correspondance telle que C et C^T soient ϵ -simpliciales entre \mathbb{K} et \mathbb{L} . Alors $H(C)$ et $H(C^T)$ forment un ϵ -entrelacement entre les modules de persistance $H_*(\mathbb{K})$ et $H_*(\mathbb{L})$.

Démonstration. Il suffit de remarquer qu'alors les applications id_X et id_Y sont subordonnées à $C^T \circ C$ et $C \circ C^T$. Par la proposition 3.4.5, on a donc $H(C) \circ H(C^T) = \text{id}$ et $H(C^T) \circ H(C) = \text{id}$. Autrement dit, pour tout $t \in \mathbb{R}$, on a les diagrammes commutatifs

$$\begin{array}{ccc} H_*(K^t) & \longrightarrow & H_*(K^{t+2\epsilon}) \\ H(C)_t \downarrow & \nearrow H(C^T)_{t+\epsilon} & \\ H_*(L^{t+\epsilon}) & & \end{array} \quad \text{et} \quad \begin{array}{ccc} H_*(L^t) & \longrightarrow & H_*(L^{t+2\epsilon}) \\ H(C^T)_t \downarrow & \nearrow H(C)_{t+\epsilon} & \\ H_*(K^{t+\epsilon}) & & \end{array}$$

Ceci montre que l'on a bien affaire à un ϵ -entrelacement. \square

Exemple 3.4.8. Soient $X, Y \subset \mathbb{R}^d$ à distance de Hausdorff plus petite que ϵ . Alors la proposition précédente montre que les modules de persistance des complexes de Čech $(H_*(\check{\text{Cech}}^t(X)))_{t \geq 0}$ et $(H_*(\check{\text{Cech}}^t(Y)))_{t \geq 0}$ sont ϵ -entrelacés.

En effet, l'application multivaluée $C = \{(x, y) \in X \times Y, \|y - x\| \leq \epsilon\}$ est une correspondance. On vérifie immédiatement que C et C^T sont ϵ -simpliciales.

Remarquons ici que l'on pouvait directement montrer l'entrelacement de l'exemple précédent, en utilisant le théorème du nerf persistant et les inclusions $X^t \subset Y^{t+\epsilon} \subset X^{t+2\epsilon}$.

3.4.3 Stabilité en Gromov-Hausdorff

On considère dans cette partie deux espaces métriques compacts (X, d_X) et (Y, d_Y) . Pour toute correspondance $C : X \rightrightarrows Y$, on définit sa distorsion :

$$\text{dist}(C) = \sup \left\{ \frac{1}{2} |d_X(x, x') - d_Y(y, y')|, (x, y), (x', y') \in C \right\}.$$

Dans ce contexte, la distance de Gromov-Hausdorff entre X et Y s'écrit

$$d_{GH}(X, Y) = \inf \{ \text{dist}(C) \text{ pour } C : X \rightrightarrows Y \text{ correspondance} \}.$$

On a alors le résultat suivant :

Proposition 3.4.9. Pour tout $\epsilon > 2d_{GH}(X, Y)$, les modules de persistance des complexes de Čech $(H_*(\check{\text{Cech}}^t(X)))_{t \geq 0}$ et $(H_*(\check{\text{Cech}}^t(Y)))_{t \geq 0}$ sont ϵ -entrelacés.

Démonstration. Soit $C : X \rightrightarrows Y$ de distorsion plus petite que ϵ . Choisissons σ un simplexe de $\check{\text{Cech}}^t(X)$, et ν un sous-ensemble fini de $C(\sigma)$. Il faut montrer que ν est un simplexe de $\check{\text{Cech}}^{t+\epsilon}(Y)$.

Soit \bar{x} un point de X tel que $\bar{x} \in \bigcap_{x \in \sigma} B(x, t)$. On a que $\forall x \in \sigma, d_X(x, \bar{x}) \leq t$. Choisissons $\bar{y} \in C(x)$. En utilisant la distorsion de C , il vient que pour tout $y \in \nu$,

$$\begin{aligned} d_Y(y, \bar{y}) &\leq d_X(x, \bar{x}) + \epsilon \\ &\leq t + \epsilon. \end{aligned}$$

On a donc montré que $\bar{y} \in \bigcap_{y \in \nu} B(y, t + \epsilon)$. Autrement dit, $\nu \in \check{\text{Cech}}^{t+\epsilon}(Y)$, et C est bien ϵ -simpliciale. Par symétrie, C^T l'est aussi. Il suffit alors d'appliquer la proposition 3.4.7. \square

4 Distance à la mesure

4.1 Introduction

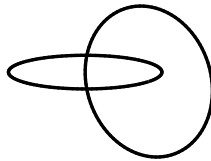
Dans les parties précédentes, nous avons vu comment estimer le diagramme de persistance d'un compact K de \mathbb{R}^d à partir d'un nuage fini de points X qui l'échantillonne.

Notons $\mathbb{V}_K = \mathbb{V}_{d_K}$ et $\mathbb{V}_X = \mathbb{V}_{d_X}$ les modules de persistance de l'homologie simpliciale des sous-niveaux des fonctions distance d_K et d_X . Si la distance de Hausdorff entre ces compacts est $d_H(K, X) = \epsilon$ alors, en vertu de la proposition 2.5.3, on a :

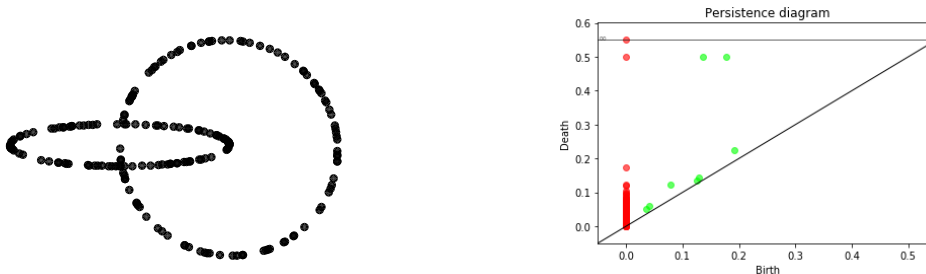
$$d_b(\mathbb{V}_K, \mathbb{V}_X) \leq \epsilon.$$

Nous avons aussi vu que le module \mathbb{V}_X est isomorphe au module de persistance construit sur l'homologie simpliciale du complexe Čech^t(X) (corollaire 3.3.5), et que nous disposons d'un algorithme pour calculer son diagramme de persistance. Quand le complexe est trop grand, on peut aussi utiliser des alternatives moins couteuses que Čech^t(X), comme le complexe de Rips.

Ceci constitue la démarche habituelle en analyse topologique des données. Considérons par exemple l'union de deux cercles plongés dans \mathbb{R}^3 :

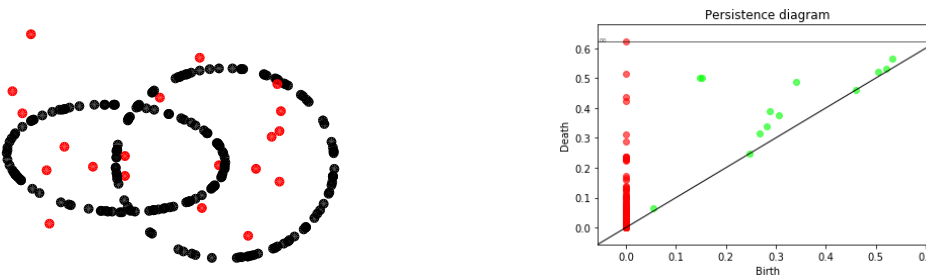


A partir d'un échantillon (ici 200 points), le complexe de Rips permet une restitution fidèle du diagramme de persistance de la fonction distance associée :



(les points de H_0 sont en rouge et ceux de H_1 en vert)

Par contre, cette méthode échoue fatalement lorsque l'échantillon comporte des points aberrants. Par exemple, nous avons ajouté ici un échantillon de 20 points uniformément distribués dans le pavé :



Les résultats précédents sont mis en défaut car la distance $d_H(K, X) = \|d_K - d_X\|_\infty$ n'est plus contrôlée.

Pour remédier à ce problème, il faut remplacer la distance d_X par une fonction qui serait robuste au bruit. C'est le dessein de la distance à la mesure, une fonction qui tient compte de la densité de la distribution de points X .

La distance à la mesure ouvre aussi une nouvelle porte : l'estimation topologique à partir de mesures.

4.2 Définition

On considère une mesure de probabilité μ sur \mathbb{R}^d , et un paramètre $m \in [0, 1]$. Soit un point $x \in \mathbb{R}^d$. Pour raisonner par analogie, rappelons que la distance de x à un compact $K \subset \mathbb{R}^d$ est le plus petit rayon $r \geq 0$ tel que la boule fermée $\overline{B}(x, r)$ intersecte K . Pour comparer x à la mesure μ , il est naturel de s'intéresser à la quantité :

$$\delta_{\mu, m}(x) = \inf\{r \geq 0, \mu(\overline{B}(x, r)) > m\}.$$

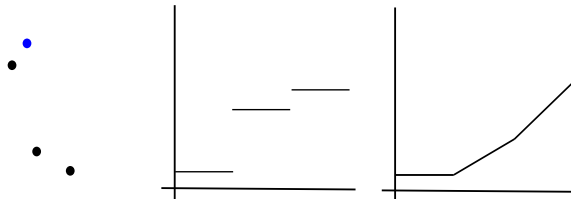
Toutefois, cette fonction n'est pas encore un bon candidat pour l'estimation topologique : elle n'est stable en fonction de μ . Par exemple, considérons les points 0 et 1 de la droite réelle, $\epsilon \in [0, 1]$, et le mélange de mesure de Dirac $\mu = \epsilon\delta_0 + (1 - \epsilon)\delta_1$. Choisissons le paramètre $m = 1/2$. Si $\epsilon > \frac{1}{2}$ la fonction $\delta_{\mu, m}$ coïncide avec $x \mapsto |x|$, tandis que si $\epsilon < \frac{1}{2}$, c'est avec $x \mapsto |1 - x|$.

La distance à la mesure μ se définit comme une moyenne quadratique des fonction $\delta_{\mu, m}$.

Définition 4.2.1. Soit $m_0 \in [0, 1]$. La distance à la mesure μ de paramètre de masse m_0 est la fonction :

$$\begin{aligned} d_{\mu, m_0} : \mathbb{R}^d &\longrightarrow \mathbb{R} \\ x &\longmapsto \sqrt{\frac{1}{m_0} \int_0^{m_0} \delta_{\mu, m}^2(x) dm} \end{aligned}$$

La figure suivante représente un point $x \in \mathbb{R}^2$ (en bleu) et une mesure μ , somme de trois mesures de Dirac (en noir). Le premier graphe est celui de la fonction $m \mapsto \delta_{\mu, m}(x)$, et le deuxième est celui de $m \mapsto d_{\mu, m}(x)$. On y observe l'effet régularisant de la définition de $\delta_{\mu, m}$.



Cas d'un nuage de points : Soit $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ un nuage fini de points, et $\mu = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. Pour tout réel $m \in]\frac{k-1}{n}, \frac{k}{n}]$, avec $k \in \mathbb{N}$, alors la fonction $\delta_{\mu, m}$ associe à chaque x sa distance avec son k -ème plus proche voisin dans X (sans s'inclure lui-même). On note p_k^x ce voisin.

Si $m_0 = \frac{k_0}{n}$, on peut écrire :

$$\begin{aligned} d_{\mu, m_0}(x) &= \frac{1}{m_0} \int_0^{m_0} \delta_{\mu, m}^2(x) dm \\ &= \frac{1}{m_0} \sum_{k=1}^{k_0} \frac{1}{n} \delta_{\mu, \frac{k}{n}}^2 = \frac{1}{k_0} \sum_{k=1}^{k_0} \|x - p_k^x\|^2. \end{aligned}$$

Le calcul de la distance à la mesure dans ce cas se réduit à une recherche de plus proches voisins.

Nous verrons dans la partie 4.4.1 une façon explicite de calculer ses sous-niveaux.

4.3 Régularité et stabilité

Cette partie est basée sur [7].

4.3.1 Lien avec les distances de Wasserstein

Nous allons établir un lien entre la distance à la mesure et un problème de transport. Soient μ et ν deux mesures sur \mathbb{R}^d de même masse. On appelle plan de transport entre μ et ν toute variable aléatoire $\pi = (X, Y)$ sur l'espace mesuré produit $\mathbb{R}^d \times \mathbb{R}^d$ telle que X soit de loi μ et Y de loi ν . On note $\Gamma(\mu, \nu)$ l'ensemble des plans de transport.

Définition 4.3.1. La 2-ème distance de Wasserstein entre μ et ν est

$$\begin{aligned} W_2(\mu, \nu) &= \inf_{(X, Y) \in \Gamma(\mu, \nu)} \sqrt{\mathbb{E}[\|X - Y\|^2]} \\ &= \inf_{\pi \in \Gamma(\mu, \nu)} \left(\int \|x - y\|^2 d\pi(x, y) \right)^{\frac{1}{2}}. \end{aligned}$$

La distance W_2 est bien définie pour les lois de probabilité admettant un moment d'ordre 2. On sait aussi qu'elle vérifie les axiomes d'une distance.

Définition 4.3.2. On dit qu'une mesure ν est une sous-mesure de μ si on a $\nu \leq \mu$ en tant que fonctions. L'ensemble des sous-mesures de μ est noté $\text{Sub}(\mu)$. Si $m_0 \in \mathbb{R}$, on notera $\text{Sub}_{m_0}(\mu)$ l'ensemble des sous-mesures de masse m_0 .

La distance à la mesure se reformule comme un problème de transport entre des sous-mesures :

Proposition 4.3.3. Pour tout $x \in \mathbb{R}^d$,

$$d_{\mu, m_0}(x) = \frac{1}{\sqrt{m_0}} \inf \{ W_2(m_0 \delta_x, \nu), \nu \in \text{Sub}_{m_0}(\mu) \},$$

et cet infimum est réalisé par une mesure $\nu \in \text{Sub}_{m_0}(\mu)$.

4.3.2 La DTM est une presque-distance

Nous étudions maintenant les propriétés de régularité de la distance à la mesure.

Proposition 4.3.4. La fonction $x \mapsto d_{\mu, m}(x)$ est 1-lipschitzienne.

Démonstration. Soit $x, y \in \mathbb{R}^d$ et $\mu_y \in \text{Sub}_m(\mu)$ la sous-mesure donnée par la proposition 4.3.3. On applique l'inégalité triangulaire :

$$\begin{aligned} d_{\mu, m}(x) &= \frac{1}{\sqrt{m}} W_2(m \delta_x, \mu_y) \\ &\leq \frac{1}{\sqrt{m}} W_2(m \delta_y, \mu_y) + \frac{1}{\sqrt{m}} W_2(\mu_x, \mu_y) \\ &= d_{\mu, m}(y) + \|x - y\|. \end{aligned}$$

On conclut en échangeant les rôles de x et y . □

On veut maintenant établir la semi-concavité de la distance à la mesure. Pour cela, nous allons utiliser la notion de sous-différentiel. Soit une fonction $f : \mathbb{R}^d \rightarrow \mathbb{R}$ et un vecteur $v \in \mathbb{R}^d$. On dit que v est dans le sous-différentiel $\partial_x f$ de f en x si

$$f(x + h) \geq f(x) + \langle h, v \rangle$$

pour tout $h \in \mathbb{R}^d$ assez petit. On montre que si f admet un sous-différentiel non-vide en tout point de \mathbb{R}^d , alors elle est convexe.

Proposition 4.3.5. La fonction $f : x \in \mathbb{R}^d \mapsto \|x\|^2 - d_{\mu,m}^2(x)$ est convexe.

Démonstration. Soit $x, y \in \mathbb{R}^d$ et $\mu_x, \mu_y \in \text{Sub}_m(\mu)$ les sous-mesures données par la proposition 4.3.3. On écrit :

$$\begin{aligned} d_{\mu,m}^2(y) &= \frac{1}{m} W_2(m\delta_y, \mu_y) = \frac{1}{m} \int \|y - h\|^2 d\mu_y(h) \\ &\leq \frac{1}{m} \int \|y - h\|^2 d\mu_x(h) \\ &= \frac{1}{m} \int \|y - x\|^2 + \|x - h\|^2 + 2\langle y - x, x - h \rangle d\mu_x(h) \\ &\leq \|y - x\|^2 + d_{\mu,m}^2(x) + \langle y - x, v \rangle, \end{aligned}$$

avec $v = \frac{2}{m} \int (x - h) d\mu_x(h)$. En utilisant $\|y - x\|^2 = \|y\|^2 - \|x\|^2 - 2\langle x, y - x \rangle$, il vient que

$$d_{\mu,m}^2(y) - \|y\|^2 \leq d_{\mu,m}^2(x) - \|x\|^2 + \langle y - x, v - 2x \rangle.$$

Ce qui prouve que $2x - v$ est dans le sous-différentiel de f en x . On conclut par caractérisation des fonction convexes par le sous-différentiel. \square

Enfin, on veut s'assurer que $d_{\mu,m}$ est propre. Puisqu'elle est continue, il suffit de montrer que pour tout $a \in \mathbb{R}^+$, la préimage $d_{\mu,m}^{-1}([0, a])$ est bornée. C'est vrai par exemple si μ est à support compact K , auquel cas on a l'inégalité :

$$d_{\mu,m}(x) \leq d_H(x, K) \leq \|x\| + d_H(K, \{0\}).$$

En rassemblant les trois propriétés que l'on vient d'énoncer, on obtient :

Corollaire 4.3.6. Si μ est à support compact et $m \in [0, 1[$, alors $d_{\mu,m}$ est une presque-distance.

Comme on l'a vu dans la partie 2.4, on a alors que le module de persistance $\mathbb{V}_{d_{\mu,m}}$ est q-moderé, et les changements de topologie correspondent aux valeurs critiques généralisées de $d_{\mu,m}$.

4.3.3 Stabilité

La distance à la mesure est stable par rapport à la mesure :

Proposition 4.3.7. Soient μ et ν deux mesures de probabilité sur \mathbb{R}^d . Alors pour tout $m \in [0, 1[$,

$$\|d_{\mu,m} - d_{\nu,m}\|_\infty \leq W_2(\mu, \nu).$$

Démonstration. Soit $x \in \mathbb{R}^d$ et $\mu_m \in \text{Sub}_m(\mu)$ donné par la proposition 4.3.3. On a $\sqrt{m}d_{\mu,m}(x) = W_2(m\delta_x, \mu_m)$. Soit π un plan de transport entre $m\delta_x$ et μ_m .

On peut alors montrer qu'il existe une sous-mesure de π qui soit un plan de transport entre μ_m et une sous-mesure ν_m de ν de masse m . Par définition, on a $W_2(m\delta_x, \mu_m) \leq W_2(m\delta_x, \nu_m)$ et $W_2(\mu_m, \nu_m) \leq W_2(\mu, \nu)$. On calcule alors :

$$\begin{aligned} d_{\mu,m}(x) &= \frac{1}{\sqrt{m}} W_2(m\delta_x, \mu_m) \leq \frac{1}{\sqrt{m}} W_2(m\delta_x, \nu_m) + \frac{1}{\sqrt{m}} W_2(\nu_m, \mu_m) \\ &\leq d_{\nu,m}(x) + \frac{1}{\sqrt{m}} W_2(\mu, \nu). \end{aligned}$$

En interchangeant les rôles de μ et ν , on obtient le résultat. \square

On finit par un résultat qui relie la distance à la mesure et la distance au support. Si μ est à support compact K , alors on voit par définition que $d_{\mu,0} = d_K$. De plus :

Proposition 4.3.8. On a une convergence uniforme

$$\|d_{\mu,m} - d_K\|_\infty \xrightarrow{m \rightarrow 0} 0.$$

4.4 Calcul et approximation

Les résultats précédents montrent que si un nuage de points $X = \{x_1, \dots, x_n\}$ échantillonne suffisamment bien la mesure μ — c'est à dire si la distance de Wassertsein entre μ et la mesure empirique $\mu^X = \frac{1}{n} \sum \delta_{x_i}$ est faible — alors les diagrammes de persistance des distances à la mesure sont proches :

$$d_b(\mathbb{V}_{d_{\mu,m}}, \mathbb{V}_{d_{\mu^X,m}}) \leq \|d_{\mu,m} - d_{\mu^X,m}\|_{\infty} \leq W_2(\mu, \nu).$$

Une première question se pose : à quelle vitesse la quantité $\|d_{\mu,m} - d_{\mu^X,m}\|_{\infty}$ tend-elle vers 0 ? Sous des hypothèses de régularité sur la mesure μ et son support, on peut obtenir des taux explicites de convergence (comme dans [13]). Mais nous n'en parlerons pas ici.

A vrai dire, le calcul algorithmique du diagramme de $\mathbb{V}_{d_{\mu^X,m}}$ n'est pas tout à fait résolu : à quoi ressemblent les sous-niveau de la fonction $d_{\mu^X,m}$? On aimerait pouvoir les écrire comme de bons recouvrements, comme ce que l'on a fait pour les fonctions distances.

Les résultats de cette partie proviennent de [14].

4.4.1 Décomposition barycentrique

On considère toujours $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$. On notera $\mu = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ sa mesure de probabilité empirique. Fixons un paramètre $m \in [0, 1[$ s'écrivant $m = \frac{k}{n}$, $k \in \mathbb{N}$.

Pour tout point $x \in \mathbb{R}^d$ et $i \leq n$, on note p_i^x sont i -ème plus proche voisin. On notera aussi $\text{bar}(x)$ le barycentre de ses k plus proches voisins :

$$\text{bar}(x) = \text{isobar}(p_1^x, \dots, p_k^x) = \frac{1}{k} \sum_{i=1}^k p_i^x.$$

On peut faire intervenir ces barycentres dans l'expression de la distance à la mesure :

$$\begin{aligned} d_{\mu,m_0}^2(x) &= \frac{1}{k} \sum_{i=1}^k \|x - p_i^x\|^2 \\ &= \frac{1}{k} \sum_{i=1}^k [\|x - \text{bar}(x)\|^2 + \|\text{bar}(x) - p_i^x\|^2 + 2\langle x - \text{bar}(x), \text{bar}(x) - p_i^x \rangle] \\ &= \|x - \text{bar}(x)\|^2 + \frac{1}{k} \sum_{i=1}^k \|\text{bar}(x) - p_i^x\|^2 + 2\langle x - \text{bar}(x), \text{bar}(x) - \frac{1}{k} \sum_{i=1}^k p_i^x \rangle \\ &= \|x - \text{bar}(x)\|^2 + \underbrace{\frac{1}{k} \sum_{i=1}^k \|\text{bar}(x) - p_i^x\|^2}_{\omega^2(x)}. \end{aligned}$$

Interprétons ce résultat. L'ensemble des k -barycentres de points de X est un ensemble fini, que nous notons $\text{Bar}_k(X)$. A chacun de ces barycentres $b \in \text{Bar}_k(X)$ est associé un poids $\omega^2(b) = \frac{1}{k} \sum_{i=1}^k \|\text{bar}(x) - p_i^x\|^2$.

Ensuite, à chaque point $x \in \mathbb{R}^d$ est associé une valeur, $\|x - \text{bar}(x)\|^2$. La distance à la mesure en x est alors la somme de ces deux contributions.

On peut vérifier que la distance à la mesure répond au problème de minimisation suivant :

Proposition 4.4.1. Pour tout $x \in \mathbb{R}^d$,

$$d_{\mu,m_0}^2(x) = \min_{b \in \text{Bar}_k(X)} \|x - b\|^2 + \omega^2(b).$$

Démonstration. Le calcul précédent montre déjà qu'on a l'inégalité " \geq ". Réciproquement, si $b \in \text{Bar}_k(X)$ est le barycentre de p_1, \dots, p_k , on a

$$\begin{aligned} \|x - b\|^2 + \omega^2(b) &= \frac{1}{k} \sum_{i=1}^k [\|x - b\|^2 + \|b - q_i\|^2] \\ &\geq \frac{1}{k} \sum_{i=1}^k (\|x - b\| + \|b - q_i\|)^2 \\ &\geq \frac{1}{k} \sum_{i=1}^k \|x - q_i\|^2 \geq \frac{1}{k} \sum_{i=1}^k \|x - p_i^x\|^2 = d_{\mu,m}^2(x). \end{aligned}$$

□

Remarque 4.4.2. Une telle fonction s'appelle une *power distance* pour la fonction de poids ω . Son calcul est relié à celui d'un diagramme de Voronoï pondéré d'ordre k , et il existe des algorithmes qui peuvent s'en charger.

Nous savons maintenant décrire les sous-niveaux de la fonction $d_{\mu,m}$: pour tout $t \geq 0$, un point x vérifie $d_{\mu,m} \leq t$ si et seulement si il existe un $b \in \text{Bar}_k(X)$ tel que $\|x - b\|^2 + \omega^2(b) \leq t^2$, c'est à dire si :

$$\|x - b\|^2 \leq t^2 - \omega^2(b).$$

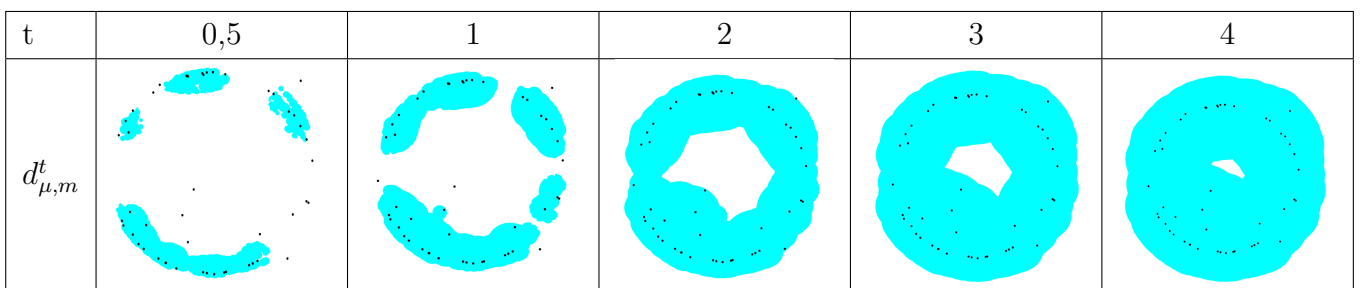
Autrement dit, $d_{\mu,m}^t$ est l'union des boules $B(b, \sqrt{t^2 - \omega^2(b)})$, où b parcourt $\text{Bar}_k(X)$. On note ce recouvrement :

$$\mathcal{U} = \{B(b, \sqrt{t^2 - \omega^2(b)}), b \in \text{Bar}_k(X)\}.$$

En conséquence, et en appliquant le théorème du nerf persistant, le diagramme de persistance $\mathbb{V}_{d_{\mu,m}}$ se calcule en considérant l'homologie simpliciale persistante du nerf de \mathcal{U} .

Tel quel, le calcul de ce recouvrement est toutefois problématique : l'ensemble des k -barycentres est de taille $\binom{n}{k}$. Cependant, certains de ces points ne sont pas utiles, car leur boule $B(b, \sqrt{t^2 - \omega^2(b)})$ est déjà contenue dans d'autres boules de \mathcal{U} . Cette remarque est à l'origine de certaines méthodes d'approximation des sous-niveaux.

Voici quelques sous-niveaux de $d_{\mu,m}$ pour $m = 0, 1$ et μ la mesure empirique d'un nuage de points bruité qui échantillonne le cercle :



4.4.2 Complexe de Čech pondéré

Plutôt que de poser sur chaque point de $\text{Bar}_k(X)$ une boule de rayon $\sqrt{t^2 - \omega^2(b)}$, comme le fait la distance à la mesure, on pourrait préférer les poser sur les points du nuage X lui-même. Cela conduit à l'approximation suivante :

Définition 4.4.3. La distance à la mesure approximée est la fonction :

$$d_{\mu,m_0}^X : z \in \mathbb{R}^d \mapsto \min_{x \in X} \sqrt{\|x - z\|^2 + d_{\mu,m}^2(x)}.$$

Comme précédemment, les sous-niveaux de $d_{\mu,m}^X$ sont des unions de boules : pour tout $t \geq 0$,

$$(d_{\mu,m}^X)^t = \bigcup_{x \in X} B(x, \sqrt{t^2 - d_{\mu,m}^2(x)})$$

Le nerf de ce recouvrement est appelé complexe de Čech pondéré, et on le note $\check{\text{Cech}}_m^t(X)$. Le module de persistance correspondant, $(H_*(\check{\text{Cech}}_m^t(X)))_{t \geq 0}$, est isomorphe à $\mathbb{V}_{d_{\mu,m}^X}$.

Voici la qualité de cette approximation :

Proposition 4.4.4. Pour tout $x \in \mathbb{R}^d$, on a les inégalités :

$$\frac{1}{\sqrt{2}}d_{\mu,m}(x) \leq d_{\mu,m}^X(x) \leq \sqrt{3}d_{\mu,m}(x).$$

Ces relations sont optimales.

On déduit de cette proposition que :

$$\log(d_{\mu,m}) \leq \log(\sqrt{3}) + \log(d_{\mu,m}^X) \text{ et}$$

$$\log(d_{\mu,m}^X) \leq \log(\sqrt{3}) + \log(d_{\mu,m}).$$

Autrement dit, les sous-niveaux de $\log(d_{\mu,m})$ et $\log(d_{\mu,m}^X)$ sont $\log(\sqrt{3})$ -entrelacés. Cela se traduit par une majoration de la distance d'entrelacement en échelle logarithmique entre les modules de persistance (définition 3.3.9) :

$$d_i^{\log}(\mathbb{V}_{d_{\mu,m}}, \mathbb{V}_{d_{\mu,m}^X}) \leq \log(\sqrt{3}).$$

Appliquons à ce résultat une dernière approximation : si le complexe $\check{\text{Cech}}_m^t(X)$ est trop coûteux à calculer, alors on peut se contenter du complexe de Rips pondéré $\text{Rips}_m^t(X)$, qui est défini comme son complexe de clique (comme dans la partie 3.3.2). Nous avons alors des inclusions entre complexes simpliciaux :

$$\check{\text{Cech}}_m^{\frac{t}{2}}(X) \subset \text{Rips}_m^t(X) \subset \check{\text{Cech}}_m^t(X).$$

Cela se traduit aussi au niveau homologique par un entrelacement logarithmique :

$$d_i^{\log}(H_*(\check{\text{Cech}}_m^t(X)), H_*(\text{Rips}_m^t(X))) \leq \log(2).$$

On vient de montrer :

Théorème 4.4.5. On a un $\log(2\sqrt{3})$ -entrelacement logarithmique entre $\mathbb{V}_{d_{\mu,m}}$, le module des sous-niveaux de la fonction distance à la mesure empirique de X , et $H_*(\text{Rips}_m^t(X))$, le module du complexe de Rips pondéré de X . Autrement dit,

$$d_i^{\log}(\mathbb{V}_{d_{\mu,m}}, H_*(\text{Rips}_m^t(X))) \leq \log(2\sqrt{3}).$$

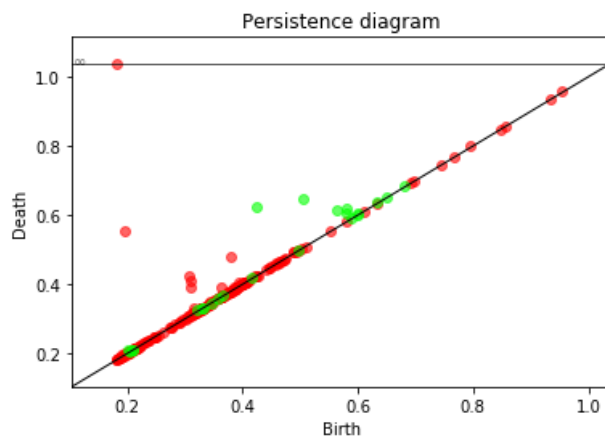
Nous mettons cela à l'épreuve dans la prochaine partie.

4.5 Simulation

Reprenons le même nuage de points X que dans l'introduction, qui échantillonne deux cercles plongés dans du bruit. On applique la méthode du complexe de Rips pondéré :

- μ est la mesure empirique de X ,
- $d_{\mu,m}$ sa distance à la mesure pour un paramètre $m \in [0, 1[$,
- $d_{\mu,m}^X$ est la distance à la mesure approximée,
- $\check{Cech}_m^t(X)$ est le complexe de Čech pondéré par $d_{\mu,m}^X$,
- $Rips_m^t(X)$ le complexe de Rips correspondant,
- et $(H_*(Rips_m^t(X)))_{t \geq 0}$ est le module de persistance de l'homologie simpliciale de ce complexe de Rips.

Nous pouvons calculer le diagramme de $(H_*(Rips_m^t(X)))_{t \geq 0}$ grâce à l'algorithme de persistance. Pour le paramètre $m = 0.1$, nous obtenons :



Sur ce diagramme, on distingue deux points rouges et deux points verts éloignés de la diagonale : les propriétés topologiques de l'espace initial ont été retrouvées.

5 D'autres structures pour l'inférence topologique

5.1 Le problème de la stabilité

Nous avons vu que l'estimation de la persistance nécessite de développer des méthodes robustes aux points aberrants.

Dans cette partie, nous considérons une fonction continue $f : \mathbb{R}^d \rightarrow \mathbb{R}^+$, et $\mathbb{V}_f = (V_t, v_s^t)$ le module de persistance de l'homologie singulière sous-niveaux $f^t = f^{-1}(]-\infty, t])$ de f :

$$H_*(f^r) \rightarrow H_*(f^s) \rightarrow H_*(f^t).$$

Une hypothèse suffisante de régularité pour f est de supposer qu'elle soit propre, auquel cas le module \mathbb{V}_f est q -modéré, et possède un diagramme de persistance bien défini (proposition 2.5.2).

En pratique toutefois, nous n'avons pas accès à la fonction f , mais seulement à une approximation \hat{f} . Le problème auquel nous sommes confrontés se formule ainsi :

Créer un module de persistance $\hat{\mathbb{V}}$, construit à partir de \hat{f} , qui soit un bon estimateur de \mathbb{V}_f (en distance bottleneck par exemple).

Exemple avec la distance à la mesure : Dans ce cadre, on considère une mesure de probabilité μ sur \mathbb{R}^d et un paramètre $m = \frac{k}{n} \in [0, 1]$. La fonction f correspond alors à :

$$f = d_{\mu, m}.$$

On dispose ensuite d'un nuage de n points X , que l'on espère suffisamment bien échantillonnés selon μ . Si on note $\hat{\mu}$ la loi empirique de ce nuage de points, alors la fonction avec laquelle on travaille est en fait :

$$\hat{f} = d_{\hat{\mu}, m}.$$

Si $\mathbb{V}_{\hat{f}}$ est le module de persistance des sous-niveaux de \hat{f} , alors on peut appliquer directement les résultats de convergence de la distance à la mesure :

$$d_b(\mathbb{V}_f, \mathbb{V}_{\hat{f}}) \leq \|f - \hat{f}\|_{\infty} \leq \frac{1}{\sqrt{m}} W_2(\mu, \hat{\mu}).$$

Le problème n'est pas résolu pour autant, puisque les sous-niveaux de \hat{f} ne sont pas calculables en pratique — c'est a priori l'union de $\binom{n}{k}$ boules (partie 4.4.1). On préfère l'approche suivante :

Créer un complexe simplicial filtré $C^t(X, \hat{f})$, d'ensemble de sommets X et dépendant de \hat{f} , tel que son module de persistance $H_*(C^t(X, \hat{f}))$ soit un bon estimateur de \mathbb{V}_f .

Ceci est le problème de l'approximation. Nous avons déjà donné un exemple d'une telle construction dans la partie 4.4.2.

Dans la suite, nous ne nous intéresserons pas à la dépendance entre X et \hat{f} (comme c'est le cas pour la distance à la mesure). Voici les propriétés de stabilité que l'on espère obtenir :

La variation en distance bottleneck de $H_*(C^t(X, \hat{f}))$ est contrôlée si l'on perturbe \hat{f} , et aussi si l'on modifie X .

5.2 Structure \mathbb{W}

5.2.1 Définition

On considère $f : \mathbb{R}^d \rightarrow \mathbb{R}^+$ une application propre et 1-lipschitzienne. Soit \mathbb{V}_f comme le module de persistance des sous-niveaux de f .

Soit $X \subset \mathbb{R}^d$. Pour $x \in X$ et $t \in \mathbb{R}$, on définit

$$r_x(t) = \begin{cases} -\infty & \text{si } t < f(x) \\ t - f(x) & \text{sinon} \end{cases},$$

de sorte à avoir des boules

$$\bar{B}(x, r_x(t)) = \begin{cases} \emptyset & \text{si } t < f(x) \\ \{x\} & \text{si } t = f(x) \\ \bar{B}(x, t - f(x)) & \text{si } t > f(x) \end{cases}.$$

Pour $t \in \mathbb{R}$, on considère alors le sous-ensemble

$$A_t = \bigcup_{x \in X} \bar{B}(x, r_x(t)) \subset \mathbb{R}^d.$$

Il est recouvert par $\mathcal{U}^t = \{B(x, r_x(t))\}_{x \in X}$. C'est un bon recouvrement (fermé), et on note \mathcal{N}^t son nerf. Enfin, on définit le module de persistance

$$\mathbb{W}_{X,f} = H_*(\mathcal{N}^t).$$

Par le théorème du nerf persistant, on peut considérer $\mathbb{W}_{X,f}$ comme le module de persistance de l'homologie singulière de A_t , ou comme celui de l'homologie simpliciale de \mathcal{N} (ils sont isomorphes).

Proposition 5.2.1. Si $X = \mathbb{R}^d$, \mathbb{V}_f et $\mathbb{W}_{X,f}$ sont isomorphes.

Démonstration. Pour tout réel $t \in \mathbb{R}$, on a $f^t \subset A_t$ par définition. Soit $y \in A_t$. Alors il existe $x \in \mathbb{R}^d$ tel que $y \in B(x, r_x(t))$, c'est à dire que $\|y - x\| \leq r_x^t = t - f(x)$. Par lipschitzianité de f , on obtient $\|f(y) - f(x)\| \leq t - f(x)$, puis $f(y) \leq t$. Donc $y \in f^t$.

On a montré l'égalité $f^t = A_t$, ce qui conclut. \square

Pour étudier la stabilité de $\mathbb{W}_{X,f}$, nous allons effectuer des contractions d'arêtes.

5.2.2 Contraction d'arêtes

Relation de domination. Soit un complexe simplicial X d'ensemble de sommets V . Soient $u, v \in V$ deux sommets. On dit que v domine u si tout simplexe maximal qui contient u contient aussi v . Autrement dit, pour tout $v_1, \dots, v_n \in V$, si $[u, v_1, \dots, v_n]$ est un simplexe de X , alors $[v, u, v_1, \dots, v_n]$ l'est aussi.

Soit $X \setminus u$ le complexe simplicial obtenu à partir de X en retirant le sommet u des simplexes. Son ensemble de sommets est $V \setminus \{u\}$. On dit qu'on a contracté l'arête $[u, v]$.

Lemme 5.2.2. Si v domine u , alors l'inclusion entre complexes simpliciaux $i : X \setminus u \hookrightarrow X$ induit un isomorphisme $H(X \setminus u) \xrightarrow{\sim} H(X)$ entre les groupes d'homologie simpliciale.

Démonstration. Soit l'application

$$\begin{aligned} j : V &\longrightarrow V \setminus \{u\} \\ u &\longmapsto v \\ x &\longmapsto x \text{ si } x \neq u \end{aligned}$$

La composée $j \circ i : V \setminus \{u\} \rightarrow V \setminus \{u\}$ coïncide avec l'application identité, et il suffit donc de montrer que j est simpliciale pour obtenir le résultat (proposition A.3.4).

Choisissons $\omega = [v_1, \dots, v_n]$ un simplexe de X . Il faut montrer que $j(\omega)$ est un simplexe de $X \setminus u$. Si ω ne contient pas u , on a $j(\omega) = \omega$. Sinon, on peut supposer que $v_1 = u$, et on écrit alors

$$j(\omega) = j([u, v_2, \dots, v_n]) = [v, v_1, \dots, v_n],$$

qui est un simplexe de X (et donc de $X \setminus u$) par domination de u par v . \square

Contraction d'arête au niveau persistance. On considère maintenant que le complexe simplicial est filtré par $f : X \rightarrow \mathbb{R}$. Pour $t \in \mathbb{R}$, on note X^t le complexe simplicial $f^{-1}((-\infty, t])$.

Soient $\epsilon \geq 0$ et u, v deux sommets de V . On dit que v ϵ -domine u si :

$$\forall t \in \mathbb{R}, \forall v_1, \dots, v_n \in V, [u, v_1, \dots, v_n] \in X^t \implies [v, u, v_1, \dots, v_n] \in X^{t+\epsilon}.$$

On note \mathbb{V}_X le module de persistance associé à (X, f) , et $\mathbb{V}_{X \setminus u}$ celui associé à $(X \setminus u, f)$.

Lemme 5.2.3. Si v ϵ -domine u , alors les applications $(i_t : X^t \setminus u \rightarrow X^{t+\epsilon})_{t \in \mathbb{R}}$ et $(j_t : X^t \rightarrow X^{t+\epsilon} \setminus u)_{t \in \mathbb{R}}$ (définies dans le lemme précédent) induisent un ϵ -entrelacement entre \mathbb{V}_X et $\mathbb{V}_{X \setminus u}$.

Démonstration. Les familles d'applications (i_t) et (j_t) sont bien des ϵ -morphisms. De plus, comme dans le lemme précédent, les applications j_t sont simpliciales, et on a alors, pour tout $t \in \mathbb{R}$, des diagrammes commutatifs :

$$\begin{array}{ccc} H(X^t \setminus u) & \xrightarrow{\quad} & H(X^{t+2\epsilon} \setminus u) \\ & \searrow i_t \quad \nearrow j_{t+\epsilon} & \\ & H(X^{t+\epsilon}) & \end{array} \quad \text{et} \quad \begin{array}{ccc} H(X^t) & \xrightarrow{\quad} & H(X^{t+2\epsilon}) \\ & \searrow j_t \quad \nearrow i_{t+\epsilon} & \\ & H(X^{t+\epsilon} \setminus u) & \end{array},$$

ce qui signifie bien que les deux modules de persistance sont ϵ -entrelacés. \square

En fait, on peut obtenir une version plus forte du résultat :

Lemme 5.2.4. Soient $\{u_i\}_{i \in I}$ et $\{v_i\}_{i \in I}$ deux familles de sommets disjointes (éventuellement infinies) telles que $\forall i \in I$, v_i ϵ -domine u_i . Alors \mathbb{V}_X et $\mathbb{V}_{X \setminus \{u_i\}_{i \in I}}$ sont ϵ -entrelacés.

Démonstration. Il suffit de considérer l'application :

$$\begin{aligned} j : V &\longrightarrow V \setminus \{u_i\}_{i \in I} \\ u_i &\longmapsto v_i \\ x &\longmapsto x \text{ si } x \neq \{u_i\}_{i \in I} \end{aligned}$$

et de raisonner comme précédemment. \square

5.2.3 Approximation et stabilité

Nous pouvons maintenant énoncer un résultat de stabilité. Pour cela, on suppose qu'il existe un compact $K \subset \mathbb{R}^d$ et une constante $c \in \mathbb{R}$ tels que :

- $f|_K \leq c$, et
- $\forall y \in \mathbb{R}^d, f(y) \geq d(y, K)$ (la distance euclidienne).

Proposition 5.2.5. Soit $X \subset \mathbb{R}^d$ qui contient K . Sous les hypothèses précédentes, les modules $\mathbb{W}_{X,f}$ et $\mathbb{W}_{K,f}$ sont c -entrelacés.

Démonstration. On note toujours \mathcal{N}^t le nerf du recouvrement $\{B(x, r_x(t))\}_{x \in X}$.

Pour tout $y \in X \setminus K$, nous allons définir un sommet $z \in K$ qui c -domine y dans le complexe simplicial filtré $(\mathcal{N}^t)_t$. On obtient alors le résultat en appliquant le lemme 5.2.4.

Choisissons $y \in X \setminus K$, et définissons $z \in \operatorname{argmin}_{x \in K} \{d(z, x)\}$ une projection de y sur K . Soit $t \geq f(y)$, et $\omega = [y, x_1, \dots, x_n]$ un simplexe de N_t qui contient y . On veut montrer que $\omega' = [z, y, x_1, \dots, x_n]$ est un simplexe de N_{t+c} .

Puisque N_t est un nerf, on a que

$$\bigcap_{i=1}^n B(x_i, r_{x_i}(t)) \cap B(y, r_y(t)) \neq \emptyset.$$

Si l'on montre que $B(y, r_y(t)) \subset B(z, r_z(t+c))$, on aura alors :

$$\bigcap_{i=1}^n B(x_i, r_{x_i}(t+c)) \cap B(y, r_y(t+c)) \cap B(z, r_z(t+c)) \neq \emptyset,$$

et le résultat sera démontré.

Or $B(y, r_y(t)) \subset B(z, r_z(t+c))$ est équivalent à $\|z - y\| + r_y(t) \leq r_z(t+c)$.

Par hypothèse, on a $f(y) \geq \|z - y\|$, et $f(z) \leq c$. Donc $r_y(t) + \|z - y\| = t - f(y) + \|z - y\| \leq t$, et $t \leq t + c - f(z) = r_z(t+c)$. \square

Proposition 5.2.6. On suppose que f est 1-lipschitzienne. Soient deux compacts X et $Y \subset \mathbb{R}^d$ tels que $d_H(X, Y) \leq \epsilon$ (distance de Hausdorff). Alors $\mathbb{W}_{X,f}$ et $\mathbb{W}_{Y,f}$ sont 2ϵ -entrelacés.

Démonstration. Cela vient des inclusions $A_X^t \subset A_Y^{t+\epsilon} \subset A_Y^{t+2\epsilon}$, et du fait que l'on peut considérer les modules de persistance $\mathbb{W}_{X,f}$ et $\mathbb{W}_{Y,f}$ comme induits par l'homologie singulière des espaces $(A_X^t)_t$ et $(A_Y^t)_t$. \square

Ces deux résultats donnent lieu à :

Corollaire 5.2.7. Soit $X \subset \mathbb{R}^d$ un ensemble qui contienne un sous-ensemble à distance de Hausdorff au plus ϵ de K . Autrement dit, on demande que la distance de Hausdorff non-symétrique

$$d_H(X; K) = \inf\{r \geq 0, K \subset X^r\}$$

soit plus petite que ϵ . Alors les modules de persistance $\mathbb{W}_{X,f}$ et \mathbb{V}_f sont $(c + \epsilon)$ -entrelacés.

Nous avons ici un résultat de robustesse face aux points aberrants : si $d_H(X; K) \leq \epsilon$ est vérifiée, alors nous pouvons rajouter à X autant de points que voulu sans modifier l'approximation $d_b(\mathbb{W}_{X,f}, \mathbb{V}_f) \leq c + \epsilon$.

5.3 Structure \mathbb{F}

5.3.1 Définition

Nous proposons maintenant une autre construction.

On a toujours $f : \mathbb{R}^d \rightarrow \mathbb{R}^+$ propre et 1-lipschitzienne, et $X \subset \mathbb{R}^d$. On définit $S_{X,f}$ comme étant le complexe simplicial filtré d'ensemble de sommets X tel que :

$$\forall \omega = [x_0, \dots, x_n] \subset X, \omega \in S_{X,f}^t \text{ si et seulement si } \max f(\text{conv}(\{x_0, \dots, x_n\})) \leq t,$$

où $\text{conv}(\{x_0, \dots, x_n\})$ signifie enveloppe convexe. Autrement dit, on ajoute un sous ensemble de X au complexe simplicial si la fonction f est assez petite sur l'enveloppe convexe de ses sommets.

On définit ensuite $\mathbb{F}_{X,f}$ comme le module de persistance de l'homologie simpliciale de $(S_{X,f}^t)_{t \in \mathbb{R}}$.

Ce complexe simplicial ne s'écrit pas naturellement comme le nerf d'un recouvrement. Pour l'étudier, nous allons utiliser les correspondances, que l'on a présenté dans la partie 3.4.2.

5.3.2 Approximation et stabilité

Cette première proposition relate la stabilité de $\mathbb{F}_{X,f}$ lorsque l'on perturbe f :

Proposition 5.3.1. Soit g une fonction telle que $\|f - g\|_\infty = \epsilon$. Alors les modules de persistance $\mathbb{F}_{X,f}$ et $\mathbb{F}_{X,g}$ sont ϵ -entrelacés.

Démonstration. Il suffit de montrer que l'on a des inclusions $S_{X,f}^t \subset S_{X,g}^{t+\epsilon} \subset S_{X,f}^{t+2\epsilon}$. Cela est clair par définition. \square

Et maintenant lorsque l'on perturbe X :

Proposition 5.3.2. Soit $Y \subset \mathbb{R}^d$ tel que $d_H(X, Y) = \epsilon$. Alors $\mathbb{F}_{X,f}$ et $\mathbb{F}_{Y,f}$ sont ϵ -entrelacés.

Démonstration. On considère l'application multivaluée :

$$C : x \in X \mapsto B(x, \epsilon) \cap Y.$$

Puisque $d_H(X, Y) = \epsilon$, C est bien une correspondance. On veut montrer que C et C^T sont ϵ -simpliciales. Soit $\omega = [x_1, \dots, x_n] \in S_{X,f}^t$. On a que $f(\text{conv}(x_1, \dots, x_n)) \leq t$. Soient $y_1 \in C(x_1), \dots, y_n \in C(x_n)$. Il faut vérifier que $f(\text{conv}(y_1, \dots, y_n)) \leq t + \epsilon$.

On utilise la 1-lipschitzianité de f : pour toute combinaison convexe $y = \sum_{i=1}^n t_i y_i$, et en notant $x = \sum_{i=1}^n t_i x_i$, on a :

$$\begin{aligned} f(x) - f(y) &\leq \|x - y\| = \left\| \sum_{i=1}^n t_i (x_i - y_i) \right\| \\ &\leq \sum_{i=1}^n t_i \|x_i - y_i\| \leq \sum_{i=1}^n t_i \epsilon = \epsilon. \end{aligned}$$

Ceci est valable pour tout $y \in \text{conv}(y_1, \dots, y_n)$, et C est bien ϵ -simpliciale.

De la même manière on montre que C^T l'est. En utilisant la proposition 3.4.7, on obtient l' ϵ -entrelacement voulu. \square

Corollaire 5.3.3. Si K est compact, alors $\mathbb{F}_{X,f}$ est q -modéré.

Démonstration. Soit $\epsilon > 0$, X un ensemble fini ϵ -proche de K en distance de Hausdorff. D'après la proposition précédente, $\mathbb{F}_{K,f}$ et $\mathbb{F}_{X,f}$ sont ϵ -entrelacés. On a donc par définition des diagrammes commutatifs en tout $t \in \mathbb{R}$:

$$\begin{array}{ccccc}
& & \xrightarrow{v_t^{t+2\epsilon}} & & \\
H_*(\mathbb{F}_{K,f}^t) & \longrightarrow & H_*(\mathbb{F}_{X,f}^{t+\epsilon}) & \longrightarrow & H_*(\mathbb{F}_{K,f}^{t+2\epsilon})
\end{array}$$

Puisque que les espaces vectoriels $H_X(\mathbb{F}_{X,f})$ sont de dimension finie (car X est fini), les applications $v_t^{t+2\epsilon}$ sont de rang fini. Ceci étant valable pour tout $\epsilon > 0$, on a montré que $\mathbb{F}_{K,f}$ est bien modéré. \square

5.3.3 Complexe de clique de \mathbb{F}

Si $S_{X,f}^t$ est le complexe simplicial qui induit le module $\mathbb{F}_{X,f}$, on définit le complexe $\text{Rips}S_{X,f}^t$ comme étant le complexe de clique de $S_{X,f}^t$.

On a vu dans la partie 3.3.2 que les modules de persistance des complexes de Čech et de Rips sont $\log(2)$ -entrelacés en échelle logarithmique.

Ce n'est pas le cas pour le module \mathbb{F} . En effet, si ω est un simplexe de $\text{Rips}S_{X,f}^t$, il se peut que ω ne soit pas dans $S_{X,f}^t$ avant un temps arbitrairement grand, comme le montre l'exemple suivant :

Exemple 5.3.4. Soit $X = \{r, rj, r\bar{j}\} \subset \mathbb{R}^2$ avec $r \leq 0$ et $1, j, \bar{j}$ les racines 3-èmes de l'unité. Soit aussi f la fonction distance au triangle X . Alors on a :

$$f(0) = d_X(0) = \left\| \frac{r + rj}{2} \right\| = \frac{r}{2}.$$

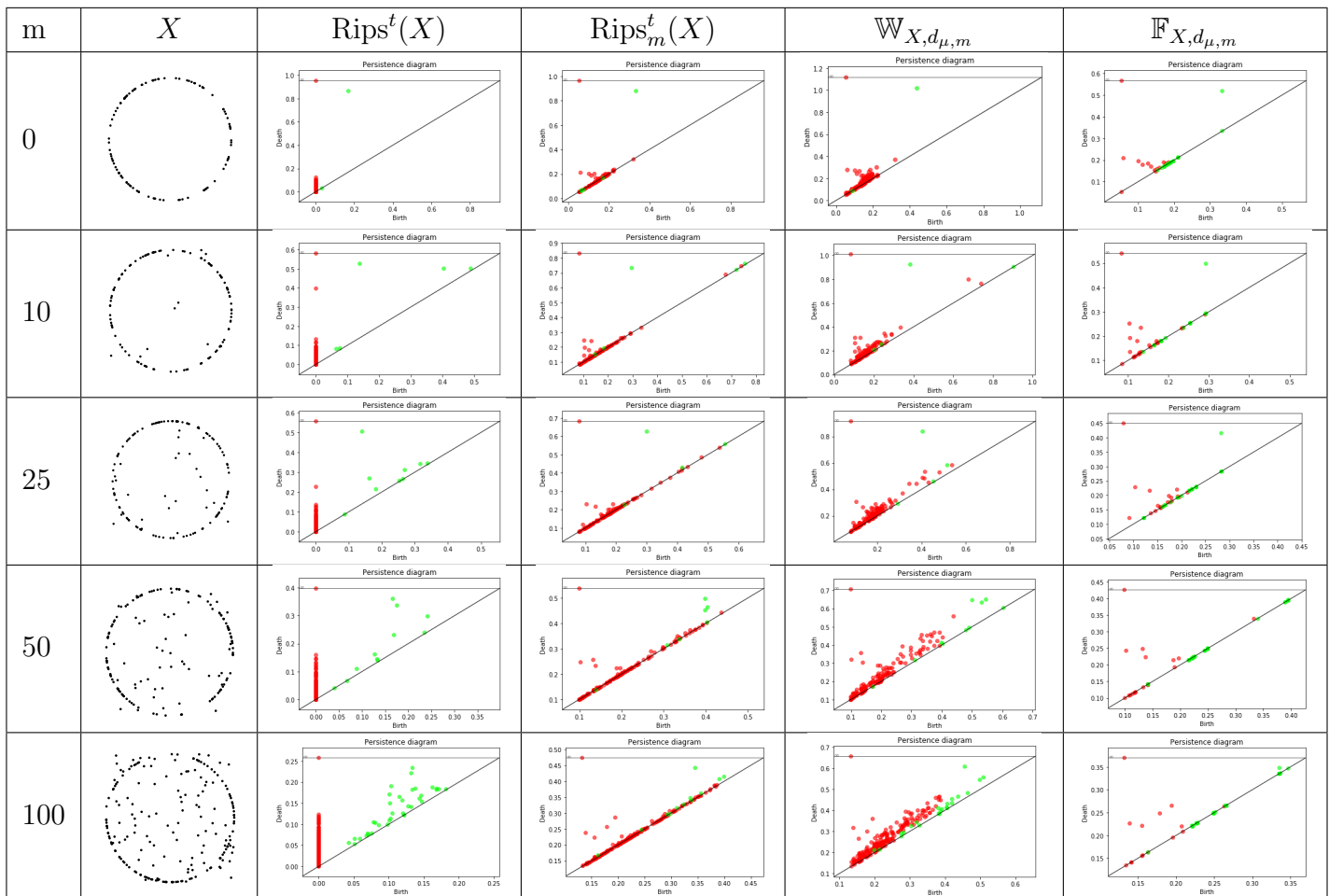
Autrement dit, le simplexe $\omega = [r, rj, r\bar{j}]$ n'apparaît pas dans $S_{X,f}^t$ avant le temps $t = \frac{r}{2}$. Pourtant, ces trois arêtes sont présentes dès $t = 0$, et donc $\omega \in \text{Rips}S_{X,f}^t$ en $t = 0$.

5.4 Simulations

Nous allons maintenant comparer les quatre structures que nous avons défini :

- le complexe $\text{Rips}^t(X)$ (partie 3.3.2),
- le complexe de Rips pondéré $\text{Rips}_m^t(X)$ (partie 4.4.2),
- le complexe de clique de $\mathbb{W}_{X,d_{\mu,m}}$ (partie 5.2), et
- le complexe de clique de $\mathbb{F}_{X,d_{\mu,m}}$ (partie 5.3).

Nous utilisons comme ensemble d'observations, noté X , l'union d'un échantillon uniforme de 100 points sur le cercle \mathbb{S}^1 et d'un échantillon uniforme de m points dans le carré $[-1, 1]^2$. Le tableau suivant rassemble les diagrammes de persistance des complexes que l'on a défini, pour différents ensembles X , et pour le paramètre $m = 0, 1$.



Nous pouvons observer que dès $m = 50$, le diagramme de $\text{Rips}^t(X)$ n'est plus pertinent. Toutefois, pour $m = 100$, on retrouve des informations topologiques saillantes dans les diagrammes de $\text{Rips}_m^t(X)$ et $\mathbb{W}_{X,d_{\mu,m}}$.

Enfin, les deux derniers diagrammes de $\mathbb{F}_{X,d_{\mu,m}}$ sont éloignés du diagramme à estimer. Cela peut signifier que la méthode consistant à prendre son complexe de clique est trop grossière.

A Formulaire d'homologie dans $\mathbb{Z}/2\mathbb{Z}$

Cette annexe se base sur [15] et [10]. Bien que la théorie de l'homologie se définisse dans un anneau commutatif quelconque, nous n'aurons ici besoin de la considérer que sur le corps $k = \mathbb{Z}/2\mathbb{Z}$.

Dans la première partie, nous définissons les groupes d'homologie à partir des complexes de chaînes. Dans les deux suivantes, nous appliquons ces définitions à deux contextes : les espaces topologiques, et les complexes simpliciaux.

A.1 Groupes d'homologie

On appelle complexe de chaînes une suite d'espaces vectoriels $C = (C_n)_{n \in \mathbb{N}}$ sur k , munie d'une suite d'applications linéaires $\delta = (\partial_{n+1} : C_{n+1} \rightarrow C_n)$ telles que :

$$\forall n \in \mathbb{N}, \partial_n \circ \partial_{n+1} = 0.$$

Ces applications sont appelées les opérateurs de bord. On définit aussi par convention $C_0 = 0$ et $\partial_0 = 0$.

A partir d'un tel complexe de chaînes, on peut définir, pour tout $n \in \mathbb{N}$:

- l'espace des n -cycles : $Z_n(C) = \ker(\partial : C_n \rightarrow C_{n-1})$,
- l'espace des n -bords : $B_n(C) = \text{im}(\partial : C_{n+1} \rightarrow C_n)$.

Enfin, le n -ème groupe d'homologie est l'espace vectoriel quotient

$$H_n(C) = Z_n(C) / B_n(C).$$

Morphisme de complexe de chaîne. Si C et D sont deux complexes de chaînes, on appelle morphisme de complexes de chaînes entre C et D toute famille de morphismes $\phi = (\phi_n : C_n \rightarrow D_n)_{n \in \mathbb{N}}$ telle que :

$$\forall n \in \mathbb{N}, \phi_n \circ \partial_{n+1} = \partial_n \circ \phi_{n+1}.$$

Un morphisme de complexes de chaînes envoie cycles sur cycles et bords sur bords, et induit donc un morphisme d'espaces vectoriels

$$\phi : H_n(C) \rightarrow H_n(D).$$

Homotopie. Soient $\phi, \psi : C \rightarrow D$ deux morphismes de complexes de chaînes. On appelle homotopie entre ϕ et ψ une suite de morphismes $(K_n : C_n \rightarrow D_{n+1})_{n \in \mathbb{N}}$ telle que :

$$\forall n \in \mathbb{N}, \phi_n - \psi_n = \partial_{n+1} \circ K_n - K_{n-1} \circ \partial_{n+1},$$

où l'on pose $K_{-1} = 0$.

Proposition A.1.1. S'il existe une homotopie entre ϕ et ψ , alors ces deux morphismes de complexes de chaînes induisent les mêmes applications entre les groupes d'homologie.

A.2 Homologie singulière

p -simplexe standard. Soit $p \in \mathbb{N}$, et (e_0, \dots, e_p) la base canonique de \mathbb{R}^{p+1} . On définit le p -simplexe standard Δ_p comme l'enveloppe convexe de cette famille :

$$\begin{aligned} \Delta_p &= \text{conv}(e_0, \dots, e_p) \\ &= \{(t_0, \dots, t_p) \in \mathbb{R}^{p+1}, t_i \geq 0 \text{ et } \sum t_i = 1\} \subset \mathbb{R}^{p+1}. \end{aligned}$$

Chaînes singulières. Soit X un espace topologique. On appelle p -simplexe singulier toute application continue $\sigma : \Delta_p \rightarrow X$.

Sa i -ème face, pour $0 \leq i \leq p$, est le $(p-1)$ -simplexe singulier défini par :

$$\partial_i \sigma = \sigma(t_0, \dots, t_{i-1}, 0, t_{i+1}, \dots, t_p).$$

On note $C_p(X)$ l'espace vectoriel libre de base l'ensemble des p -simplexes singuliers. Ses éléments s'écrivent $\sum_{i \in I \text{ fini}} \epsilon_i \sigma_i$, avec $\epsilon_i \in k$.

On définit le bord d'un p -simplexe singulier σ comme :

$$\partial_p \sigma = \sum_{i=0}^p \partial_i \sigma \in C_{p-1}(X).$$

Par linéarité, cette application se prolonge en un morphisme $\partial_p : C_p(X) \rightarrow C_{p-1}(X)$.

Proposition A.2.1. On a $\partial_p \circ \partial_{p+1} = 0$.

Nous venons de construire un complexe de chaîne $(C_i(X), \partial_i)_{i \in \mathbb{N}}$. Nous pouvons alors utiliser la partie précédente pour définir ses cycles, ses bords et ses groupes d'homologie.

Proposition A.2.2. Une application continue entre deux espaces topologiques induit un morphisme entre leurs complexes de chaînes singulières.

Homotopies. Soient $f, g : X \rightarrow Y$ deux applications continues entre des espaces topologiques. Une homotopie entre f et g est une application $H : X \times [0, 1] \rightarrow Y$ telle que $H(\cdot, 0) = f$ et $H(\cdot, 1) = g$.

Proposition A.2.3. Deux applications continues homotopes induisent les mêmes morphismes entre les groupes d'homologie.

Enfin, on dit que deux espaces topologiques X et Y sont du même type d'homotopie si il existe deux applications $f : X \rightarrow Y$ et $g : Y \rightarrow X$ telles que $g \circ f$ et $f \circ g$ soient homotopes aux applications identité.

Proposition A.2.4. Si deux espaces topologiques sont du même type d'homotopie, alors leurs groupes d'homologie singulière sont isomorphes.

A.3 Homologie simpliciale

Complexe simplicial. Soit V un ensemble, appelé ensemble de sommets. Un complexe simplicial (combinatoire) sur V est un sous-ensemble $X \subset \mathcal{P}(V)$ tel que :

- $\forall \sigma \in X$, σ est fini,
- $\forall \sigma \in X, \forall \tau \subset \sigma, \tau \in X$.

Dans ce contexte, on appelle p -simplexe les éléments de X de cardinal $p+1$.

Chaînes simpliciales. Toujours avec un complexe simplicial X , on note $C_p(X)$ l'espace vectoriel libre de base l'ensemble des p -simplexes. Ses éléments s'écrivent $\sum_{i \in I \text{ fini}} \epsilon_i \sigma_i$, avec $\epsilon_i \in k$.

On définit le bord d'un p -simplexe singulier σ comme :

$$\partial_p \sigma = \sum_{\tau \subset \sigma}^p \tau \in C_{p-1}(X).$$

Par linéarité, cette application se prolonge en un morphisme $\partial_p : C_p(X) \rightarrow C_{p-1}(X)$.

Proposition A.3.1. On a $\partial_p \circ \partial_{p+1} = 0$.

Nous venons encore de construire un complexe de chaîne $(C_i(X), \partial_i)_{i \in \mathbb{N}}$, et nous pouvons donc définir ses cycles, ses bords et ses groupes d'homologie.

Applications simpliciales. Soient X et Y des complexes simpliciaux d'ensembles de sommets V et W . On dit que l'application $f : V \rightarrow W$ est simpliciale si :

$$\forall \sigma \in X, f(\sigma) \in Y.$$

Proposition A.3.2. Une application simpliciale entre deux complexes simpliciaux induit un morphisme entre leurs complexes de chaînes singulières.

Applications contigües. Deux applications simpliciales $f, g : V \rightarrow W$ sont dites contigües si :

$$\forall \sigma \in X, f(\sigma) \cup g(\sigma) \in Y.$$

Proposition A.3.3. Deux applications simpliciales contigües induisent les mêmes morphismes entre les groupes d'homologie.

Proposition A.3.4. Soient $f : V \rightarrow W$ et $g : W \rightarrow V$ deux applications simpliciales telles que $g \circ f$ et $f \circ g$ soient contigües avec les applications identité. Alors les groupes d'homologie simpliciale de X et Y sont isomorphes.

Références

- [1] Marc Glisse et Steve Oudot Frédéric Chazal, Vin de Silva. *The Structure and Stability of Persistence Modules*. SpringerBriefs in Mathematics, 2016.
- [2] Gunnar Carlsson Afra Zomorodian. Computing persistent homology. *Discrete and Computational Geometry*, 2005.
- [3] John Harer David Cohen-Steiner, Herbert Edelsbrunner. Stability of persistence diagrams. *Discrete and Computational Geometry*, 2007.
- [4] William Crawley-Boevey. Decomposition of pointwise finite-dimensional persistence modules. *arXiv :1210.0819v3*, 2014.
- [5] Michael Lesnick. The optimality of the interleaving distance on multidimensional persistence modules. *arXiv :1106.5305v1*, 2017.
- [6] Damian Mihai Michèle Audin. *Théorie de Morse et homologie de Floer*. EDP Sciences, 2010.
- [7] Quentin Mérigot. Détection de structure géométrique dans les nuages de points. <https://tel.archives-ouvertes.fr/tel-00443038v2>, 2010.
- [8] André Lieutier. Any open bounded subset of \mathbb{R}^n has the homotopy type as its medial axis. *Computer-Aided Design*, 2004.
- [9] Mariette Yvinec Jean-Daniel Boissonnat, Frédéric Chazal. *Geometric and Topological Inference*. à paraître.
- [10] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.
- [11] Steve Y. Oudot Frédéric Chazal. Towards persistence-based reconstruction in euclidean spaces. *arXiv :0712.2638*, 2007.
- [12] Steve Oudot Frédéric Chazal, Vin de Silva. Persistence stability for geometric complexes. *arXiv :1207.3885*, 2013.
- [13] Bertrand Michel Frédéric Chazal, Pascal Massart. Rates of convergence for robust geometric inference. *arXiv :1505.07602*, 2015.
- [14] Mickaël Buchet. Topological inference from measures. <https://hal.inria.fr/tel-01108521>, 2015.
- [15] Frédéric Paulin. Topologie algébrique élémentaire. 2009.