Topological Data Analysis

D'après Julien Tierny et Frédéric Chazal

2 décembre 2025

Table des matières

1 Homologie Simplicale **3**
1.1 Complexe Simpliciaux . . . 3
1.2 Homologie simpliciale . . . 4
1.2.1 Rappels d'homotopie . . . 4
1.2.2 Groupes d'homologie . . . 4

2 Persistence Topologique **6**
2.1 Invariance Topologique et Filtrations . . . 6
2.1.1 Homologie Singulière . . . 6
2.1.2 Filtrations . . . 6
2.2 Homologie Persistente . . . 7
2.2.1 Sur les fonctions . . . 7
2.2.2 Sur les filtrations . . . 8
2.3 Calcul de filtrations . . . 9
2.3.1 Complexes de Čech, de Vietoris-Rips, et autres . . . 9
2.3.2 Modules de Persistence . . . 10
2.3.3 Théorèmes de Stabilité . . . 11

3 Homologie pratique, échantillonnage et apprentissage **13**
3.1 Résilience statistique . . . 13
3.1.1 Calcul statistique . . . 13
3.1.2 Paysages de persistence et rééchantillonnage . . . 14
3.1.3 Bruit et méthode d'échantillonnage . . . 15
3.2 Persistence et apprentissage automatique . . . 17
3.2.1 Représentation de persistence . . . 17
3.2.2 Différentiabilité de la persistence . . . 18
3.2.3 Fonctions sur la persistence . . . 19
3.2.4 Densité de diagrammes de persistence espérés . . . 20

4 Fonctions de Morse **23**
4.1 Fonction Lisses 23
4.2 Représentation des fonctions de Morse 24
4.3 Comparaison Topologique 25
4.3.1 Simplifcation et Empreinte Topologique 25
4.3.2 Graphes de Reeb 26

5 Théorie de Morse Discrète **27**
5.1 Complexe de Morse 27

6 Inférence Topologique **27**

Résumé

mailto:julien.tierny@sorbonne-universite.fr mailto:frederic.chazal@inria.fr https://julien-tierny.github.io/topologicalDataAnalysisClass.html https://geometrica.saclay.inria.fr/team/Fred.Chazal/MVA2025

Introduction

Méthodes algorithmiques d'analyse topologique de données, particulièrement en science et en ingénierie.

Le but est de partir de données, sous forme de maillages et maillables, et de retrouver des structures au sein de jeux de données. Partant d'une carte (considérée comme jeu de données brutes), avec des features intéressantes, pour pouvoir raisonner sur l'espace, on passe à une représentation abstraite, par exemple comme un graphe, et c'est sur cette structure de données sous-jacente qu'on va raisonner. Ici, on peut ajouter des filtres pour redéfinir le maillage et donc redéfinir le résultat du raisonnement. Plus généralement, on veut construire une carte à partir d'un jeu de données. En astrophysique, par exemple, on modélise la croissance de l'univers à une grande échelle, on la simule par une grille de voxel, on estime la densité de matière noire sur chaque voxel, et on découvre une sorte de géométrie ressemblant à des neurones lorsqu'on trouve aussi des groupes de galaxies, formant une "toile cosmique". On peut calculer les connexions avec des complexes simpliciaux dits de Morse-Smale, dont on peut extraire une structure de graphes.

Ainsi, on extrait de la structure d'un ensemble de données, de manière robuste et indépendante de l'échelle, par comparaison et extraction de propriétés. Sous le capot, on fait :

- Homologie Simpliciale
- Théorie de Morse
- Homologie Persistente

Pour des données numériques, étant données un échantillon de points dans un espace euclidien, par exemple, on peut les représenter et objectiver des représentations géométriques apparaissant. On a des manières de mailler l'ensemble (triangulation de Delaunay, par exemple) qui amènent à des indicateurs qui nous expliquent où sont répartis les données, par exemple avec des noyaux pour estimer la densité. Avec une fonction scalaire sur un maillage, on définit une filtration, et on regarde les propriétés de la fonction, comme les optima locaux et on en extrait une structure algébrique (complexe de Morse-Smale) qui nous donne une structure algébrique. On obtient des générateurs, et des composantes "connexes".

On a ce genre de densité de pixels, par exemple la hauteur de surface de la mer qui permet de remarquer les vortexs, en chimie quantique ou des spectrogrammes d'enregistrement vocaux. On part d'un domaine géométrique et d'un signal sur ce domaine, signal qui exhibe des patternes géométriques qu'on souhaite quantifier. Ceci permet l'extraction de propriétés, la segmentation, la réduction de dimension et autres. Dans le cas de points en grande dimension, on a une unique théorie qui s'applique très généralement.

En terme de logiciels, on a le TTK (ParaView >= 5.10) et Gudhi (bibliothèque python).

1 Homologie Simplicale

Les données reçues, parfois, vont contenir explicitement la géométrie avec une construction combinatoire. On supposera qu'on aura une donnée d'entrée linéaire par morceau sur un complexe simplical.

1.1 Complexe Simpliciaux

Définition 1.1 Un *d-simplexe* est l'enveloppe convexe σ de $d+1$ points affinement indépendants dans l'espace euclidien $\mathbb{R}^n$ avec $0 \leq d \leq n$. On dit que d est la *dimension* du simplexe.

Une ligne est un 1-simplexe, un triangle un 2-simplexe et un tétrahèdre un 3-simplexe.

Définition 1.2 Une *face* τ d'un simplexe σ est un simplexe construit par un ensemble non vide des $d+1$ poits définissant σ. On note $\tau \leq \sigma$ et τ_i une face de dimension i. On dit aussi que σ est une *coface* de τ.

Selon la définition, on a $\sigma \leq \sigma$.

Définition 1.3 Un *complexe simplicial* $\mathcal{K}$ est une collection finie non-vide de simplexes $\{\sigma_i\}$, telle que :

1. $\tau \leq \sigma \Rightarrow \tau \in \mathcal{K}$;
2. $\sigma_i \cap \sigma_j$ est soit une face, soit vide.

Définition 1.4 *L'étoile* d'un simplexe $\sigma \in \mathcal{K}$ est l'ensemble des simplexes de $\mathcal{K}$ qui contiennent σ :

$$\mathrm{St}(\sigma) = \{\tau \in \mathcal{K} | \sigma \leq \tau\}$$

On note $\mathrm{St}_d(\sigma)$ les d-simplexes de $\mathrm{St}(\sigma)$.

C'est l'ensemble des cofaces de σ dans $\mathcal{K}$. C'est le plus petit voisinage combinatoire autour d'un simplexe.

Définition 1.5 Le *lien* d'un simplexe σ est l'ensemble des faces de $\mathrm{St}(\sigma)$ disjointes de σ :

$$\mathrm{Lk}(\sigma) = \{\tau \leq \sigma' \,|\, \sigma' \in \mathrm{St}(\sigma) \wedge \tau \cap \sigma = \varnothing\}$$

On définit de même le *d-lien* $\mathrm{Lk}_d(\sigma)$ en remplaçant St par St_d dans la définition

C'est en quelque sorte la bordure du voisinage combinatoire de lui-même.

En réalité on va considérer que les sommets (ou 0-simplexes) sont des points, et que les d-simplexes sont des ensembles de points. Ceci définit une notion de complexe simplicial abstrait, utile lorsqu'on n'a pas d'immersion dans un espace euclidien, ou alors dans un espace euclidien en trop grande dimension. On relâche ici la condition d'intersection, puisqu'on n'a plus de structure géométrique de l'espace.

Un exemple de complexe simplicial abstrait est le complexe de Rips ou complexe de Vietori-Rips. Étant donné un nuage de points avec une métrique :

- Le diamètre d'un ensemble P est défini par : $\varnothing(P) = \sup\{d(x,y) \mid x,y \in P\}$
- On construit un complexe simplicial $p \leq p_{max}$ de sorte que tous $p+1$ points dont le diamètre est plus petit qu'une valeur seuil d_{max}.

Le complexe de Rips est une généralisation de la notion de graphe de voisinage.

Définition 1.6 *L'espace sous-jacent* à un complexe simplicial est l'union des simplexes du complexe.

Définition 1.7 La *triangulation* $\mathcal{T}$ d'un espace topologique X est un complexe simplicial $\mathcal{K}$ dont l'espace sous-jacent $|\mathcal{K}|$ est homéomorphe à X.

Une triangulation d'un espace est donc un complexe simplicial abstrait.

Définition 1.8 Une *d-variété* M est un espace topologique dans lequel tout élément m a un voisinage ouvert homéomorphe à une d-boule euclidienne.

Définition 1.9 La triangulation d'une d-variété est appelée *d-variété linéaire par morceaux* (variété PL).

Représenter en mémoire un complexe simplicial est très couteux : il faut, pour chaque dimension, une liste des hyperarêtes (ou d-simplexes) en les représentant par un indice de sommet.

1.2 Homologie simpliciale

1.2.1 Rappels d'homotopie

Définition 1.10 Un *chemin* p dans C est un homéomorphisme d'un intervalle réel vers l'objet C. On dit que C est *connexe (par arcs)* si pour tous deux points il existe un chemin dans C les reliant.

Définition 1.11 Une *composante connexe* d'un objet est un sous-ensemble connexe (par arcs) maximal de l'objet.

Définition 1.12 Une *homotopie* entre deux fonctions continues f et g de X vers Y est une fonction continue $H : X \to [0,1] \to Y$ du produit d'un espace topologique X par l'intervalle unité vers un espace topologique Y de sorte que $H(x,0) = f(x)$ et $H(x,1) = g(x)$ pour tout $x \in X$. S'il existe une homotopie entre f et g on dit que f et g sont *homotopes*.

Définition 1.13 Si dans un espace X, tous les chemins entre tous deux points sont homotopes, on dit que X est *simplement connexe*.

Le disque est simplement connexe, mais pas le disque privé de 0.

Définition 1.14 La *caractéristique d'Euler* d'une triangulation T d'un espace topologique est la somme alternée des nombres des i-simplexes :

$$\chi(T) = \sum_{i=0}^{d} (-1)^i \, |\sigma_i|$$

Proposition 1.15 La caractéristique d'Euler est invariante par homéomorphisme.

1.2.2 Groupes d'homologie

Définition 1.16 Une *p-chaine* est une somme (formelle) de p-simplexes. On suppose que l'opérateur somme est défini modulo 2.

Ici, la somme est réellement la différence symmétrique (ou la disjonction exclusive) sur $\mathbb{F}_2^{|\sigma_p|}$, et définit le groupe C_p des p-chaînes. Le rang de C_p est $|\sigma_p|$ et son ordre est $2^{|\sigma_p|}$

On peut généraliser à des coefficients plus généraux. Informatiquement, il faut voir la notion de chaîne comme un masque binaire sur l'ensemble des p-simplexes, et l'addition comme la disjonction exclusive.

Définition 1.17 *L'opérateur de bordure* ∂ d'un p-simplexe renvoie la $(p-1)$-chaîne des $(p-1)$-faces du simplexe. On l'étend aux p-chaînes comme un morphisme de $C_p \to C_{p-1}$.

Pour un triangle, c'est l'ensemble de ses arêtes. Pour une arête, c'est l'ensemble des extrémités.

Proposition 1.18 L'opérateur de bordure définit une suite exacte appelée le complexe de chaîne associé au complexe K de dimension d :

$$\{0\} \to C_d(K) \xrightarrow{\partial} C_{d-1}(K) \xrightarrow{\partial} \cdots \xrightarrow{\partial} C_{k+1}(K) \xrightarrow{\partial} C_k(K) \xrightarrow{\partial} \cdots \xrightarrow{\partial} C_1(K) \xrightarrow{\partial} C_0(K) \xrightarrow{\partial} O$$

Définition 1.19 Un *p-cycle* est une p-chaîne dont la bordure est vide. On définit Z_p le groupe des p-cycles comme sous-groupe de C_p.

Définition 1.20 Le groupe B_p des p-bordures est l'image de $C_{(p+1)}$ par ∂.

Lemme 1.21 Pour tout $x \in C_p, p \geq 2$, $\partial\partial x = 0$.

Démonstration. Il suffit de vérifier le résultat sur les p-simplexes et d'étendre par somme. Puisque pour tout $(p-2)$-faces τ on a exactement 2 $(p-1)$-co-faces de τ dans un p-simplexe σ, on a le résultat. ■

On obtient directement :

Proposition 1.22 B_p est un sous-groupe de Z_p.

Si on a trois 1-simplexes e_1, e_2, e_3 mais pas leur coface commune $\{e_1, e_2, e_2\}$, on a un exemple d'inclusion stricte. Ceci nous amène à définir la notion de trou, en isolant les cycles :

Définition 1.23 Le *p-ème groupe d'homologie* H_p est le quotient de Z_p par B_p.

Démonstration. B_p étant un sous-groupe de Z_p, H_p est bien défini. ■

Géométriquement, on peut étendre un p-cycle à un autre p-cycle lorsqu'ils encapsulent le même "trou", c'est-à-dire lorsque qu'on peut "étendre" le premier cycle en encapsulant un $(p+1)$-simplexe. Une classe d'homologie est un élément de H_p, ou plutôt sa classe d'équivalence dans Z_p.

Pour calculer $|H_p|$, on énumère C_p, on élimine les chaînes de bordure non-vide pour calculer Z_p, et on peut ensuite énumérer les classes d'homologie.

Définition 1.24 On définit le *p-ème nombre de Betti* β_p comme le rang du groupe H_p. Ici, c'est $\log_2 |H_p|$.

La formule logarithmique pour β_p vient du calcul modulo 2 dans notre opération de groupe.

Proposition 1.25 La caractéristique d'Euler d'une triangulation T d'un espace topologique X de dimension d vérifie :

$$\chi(T) = \sum_{i=0}^{d} (-1)^i \beta_i(T)$$

On a des interprétentations des nombres de Betti en faible dimension. Par exemple, $\beta_0(K)$ est le nombre de composantes connexes de K.

2 Persistence Topologique

2.1 Invariance Topologique et Filtrations

2.1.1 Homologie Singulière

On a le résultat important suivant, qui va nous permettre de simplifier la manière de définir l'homologie, notamment dans la représentation des simplexes :

Théorème 2.1 Si K et K' sont des complexes simpliciaux de supports homéomorphes, leurs groupes d'homologie sont isomorphes et leurs nombres de Betti sont égaux.

Démonstration. « Débrouille-toi, normalien. » ■

On va donc définir l'homologie singulière, sur tout espace topologique, en considérant à homotopie près. On note Δ_k le k-simplexe standard dans $\mathbb{R}^k$.

Définition 2.2 Un *k-simplexe singulier* dans un espace topologique X est une fonction continue $\sigma : \Delta_k \to X$.

On reprends les constructions de l'homologie simpliciale pour les complexes singuliers : c'est la définition de l'homologie singulière.

Proposition 2.3 L'homologie singulière est définie pour tout espace X. Si X est homotopiquement équivalent au support d'un complexe simpliciel, les homologies singulières et simpliciales coïncident.

Proposition 2.4 Si $f : X \to Y$ est continue, et $\sigma : \Delta_k \to X$ est un simplexe dans X, alors $f \circ \sigma$ est un simplexe dans Y, et f définit donc une application linéaire entre les groupes d'homologie :

$$f_\sharp : H_k(X) \to H_k(Y)$$

Si f est un homéomorphisme ou une équivalence d'homotopie, alors $f_\sharp$ est un isomorphisme.

R En particulier, si $X \subset Y$, l'application d'inclusion induit une application linéaire d'homologie.

2.1.2 Filtrations

Définition 2.5 Un *complexe simplicial filtré (ou une filtration)* $\mathbb{K}$ construit sur un ensemble X est une famille $\{K_a \,|\, a \in T\}$, où $T \subseteq \mathbb{R}$, de sous-complexes d'un certain complexe simplicial fixé K avec ensemble de sommets X de sorte que $K_a \subseteq K_b$ pour tout $a \leq b$.

L'homologie persistente d'un complexe simplical filtré encode l'évolution de l'homologie des sous-complexes.

Définition 2.6 Une *filtration d'un complexe simplicial fini* K est une séquence de sous-complexes telle que :

1. $\varnothing = K^0 \subset K^1 \subset \cdots \subset K^m = K$
2. $K^{i+1} = K^i \cup \sigma^{i+1}$ où σ^{i+1} est un simplexe de K.

La famille des ensembles de sous-niveau pour une fonction est un exemple de filtration. On verra plus bas comment définir de telles filtrations dans plus de cas. On a toutefois un algorithme pour calculer itérativement l'homologie simpliciale d'un complexe étant donné une filtration de celui-ci :

Définition 2.7 Un $(k+1)$-simplexe σ^i est dit *positif* s'il est contenu dans un $(k+1)$-cycle de K^i. Il est dit *négatif* sinon.

Algorithme 1 Calcul d'homologie (simpliciale)

InputUne filtration $\left(K^i\right)_{i\leq m}$ d'un complexe simplical K en dimension d

$\beta_i \leftarrow 0$
for $i \in \{1,\dots,m\}$ **do**
  $k \leftarrow \dim \sigma^i - 1$
  if σ^i est dans un $(k+1)$-cycle de K^i **then**
    $\beta_{k+1} \leftarrow \beta_{k+1} + 1$
  else
    $\beta_k \leftarrow \beta_k - 1$
return $(\beta_0,\dots,\beta_d)$

Un $(k+1)$-simplexe positif crée un $(k+1)$-cycle dans K^i. Un $(k+1)$-simplexe négatif détruit un k-cycle dans K^i.

On veut donc vérifier :

$$\beta_k(K) = |k\text{-simplexes positifs}| - |(k-1)\text{-simplexes négatifs}|$$

On a :

Lemme 2.8 Si σ^i est un k-simplexe positif, il existe un unique k-cycle c_σ tel que :

1. c_σ n'est pas une bordure dans K^i
2. c_σ contient σ^i mais pas d'autre k-simplexe positif.

Démonstration. Par induction sur l'ordre d'apparition des simplexes dans la filtration. ■

Preuve de correction de l'algorithme 1. Si σ^i est contenu dans un $(k+1)$-cycle c de K^i, alors ce cycle n'est pas une bordure dans K^i. De plus, c ne peut pas être homologue à un cycle dans K^{i-1}, et donc $\beta_{k+1}\left(K^i\right) \geq \beta_{k+1}(K^{i-1}) + 1$. Si σ^i n'est contenu dans aucun $(k+1)$-cycle c de K^i, alors $\partial\sigma^i$ n'est pas une bordure dans K^{i-1} et donc $\beta_k(K^i) \leq \beta_k\left(K^{i-1}\right) - 1$. ■

Cela pose quelques questions, qui vont nous conduire à introduire l'homologie persistente.

2.2 Homologie Persistente

2.2.1 Sur les fonctions

Pour définir les diagrammes de persistence pour une fonction $f : X \to \mathbb{R}$, on étudie ses ensembles de sous-niveau. On représente, pour chaque dimension d d'homologie, la "durée de vie" d'une propriété topologique de dimension d. Ces propriétés topologiques sont, entre autres, observées par la variation du d-ème nombre de Betti. On représente alors sur un graphe 2-D, en abscisse, la valeur x pour laquelle une propriété apparaît, et en ordonnée la valeur x' pour laquelle la propriété disparaît. On a nécessairement $x' > x$. On notera $D_{f,d}$ le diagramme défini ci-dessus, comme son ensemble de points du plan $\mathbb{R}^2_+$.

On définit alors une distance sur deux tels diagrammes :

Définition 2.9 La *distance infinie de Wasserstein, ou distance du goulot* entre deux diagrammes D_1 et D_2 est définie par :

$$d_B\left(D_1, D_2\right) = \inf_{\gamma\in\Gamma} \sup_{p\in D_1} \|p - \gamma(p)\|_\infty$$

où Γ est l'ensemble des bijections entre D_1 et D_2.

On note que, les normes étant équivalentes sur $\mathbb{R}^2$, la distance de Wasserstein 2 est, à un facteur près, la distance infinie.

R Pour pouvoir obtenir des bijections, on doit souvent *augmenter* les diagrammes, en projetant les points de l'un sur la diagonale de l'autre et réciproquement.

On aura alors un théorème important de stabilité :

Théorème 2.10 Pour toutes fonctions *modérées* $f, g : \mathbb{X} \to \mathbb{R}$, $d_B(D_f, D_g) \leq \|f - g\|_\infty$.

Cette définition, avec les mains, va être précisée et étendue plus bas.

2.2.2 Sur les filtrations

On va maintenant étendre la notion de diagrammes de persistences aux filtrations de complexes simpliciaux. On a une relation fondamentale :

Proposition 2.11 Si $t \leq t'$, $f^{-1}(]-\infty, t]) \subseteq f^{-1}(]-\infty, t'])$. Si f est définie sur les sommets d'un complexe simplicial K, et étendue de sorte que

$$f(\sigma = [v_0, \ldots, v_k]) = \max f(v_i),$$

alors les ensembles de sous-niveau de f définissent une filtration du complexe simplcial K.

Il suffit maintenant d'adapter l'algorithme 1 ci-dessus pour maintenir une base d'homologie et les paires naissance-mort d'une propriété. On notera $H_k^i = H_k\left(K^i\right)$, et on va construire des bases par récurrence (les H_k sont des $\mathbb{F}_2$-espaces vectoriels).

La base de H_k^0 est vide, puisque l'ensemble est vide. Si on a construit une base de H_k^{i-1}, on a deux cas :

1. Si σ^i est un k-simplexe positif, alors on ajoute la classe d'homologie du cycle c^i associé à σ^i par le Lemme 2.8 à la base de H_k^{i-1} pour obtenir une base de H_k^i.
2. Si σ^i est un $(k+1)$-simplexe négatif :
   - On dénote $c^{j_1}, \ldots, c^{j_p}$ les cycles associés aux simplexes positifs $\sigma^{j_1}, \ldots, \sigma^{j_p}$ de la base de H_k^{i-1}.
   - On pose $d = \partial\sigma^j = \sum_{k=1}^p \varepsilon_k c^{j_k} + b$
   - On pose $l(i) = \max\{j_k \,|\, \varepsilon_k = 1\}$
   - On enlève la classe d'homologie de $c^{l(i)}$ pour obtenir une base de H_k^i.

Ceci explique comment modifier l'algorithme 1 pour calculer les diagrammes de persistence. Cependant, avant de réécrire l'algorithme, on va s'intéresser à un test algorithmique pour vérifier que σ^j est positif ou négatif. Pour ce faire, on introduit la matrice de l'opérateur de bordure. On rappelle qu'on se donne une filtration : d'un complexe simplicial fini d-dimensionnel $\varnothing = K^0 \subset K^1 \subset \cdots \subset K^m = K$ telle que $K^{i+1} = K^i \cup \sigma^{i+1}$ où σ^{i+1} est un simplexe de K.

Définition 2.12 On pose $M = (m_{i,j})_{1 \leq i,j \leq m}$ telle que $m_{i,j} = 1$ si, et seulement si σ^i est une face de σ^j et vaut 0 sinon. C'est la *matrice de l'opérateur d'inclusion.*

Pour toute colonne C_j, on définit donc $l(j)$ par :

$$(i = l(j)) \Leftrightarrow (m_{i,j} = 1 \wedge m_{i',j} = 0, \forall i' > i)$$

On obtient une version matricielle de l'algorithme de persistence :

Dans le pire des cas, on a un algorithme en $\mathcal{O}(m^3)$.

Démonstration. À chaque étape de l'algorithme, la colonne C_j représente une chaîne de la forme :

$$\partial\left(\sigma^j + \sum_{i<j} \varepsilon_i \sigma^i\right), \varepsilon_i \in \{0, 1\}$$

Algorithme 2 Algorithme de Persistence, version Matricielle

InputUne filtration $\varnothing = K^0 \subseteq \cdots \subseteq K^m = K$ d'un complexe simplical d-dimensionnel de sorte que $K^{i+1} = K^i \cup \sigma^{i+1}$ où σ^{i+1} est un simplexe de K

Calculer la matrice M de l'opérateur de bordure.
for $j \in \{0, \ldots, m\}$ **do**
  while $\exists j' < j, l(j') = l(j)$ **do**
    $C_j \leftarrow C_j + C_{j'} \mod 2$
return Paires $(l(j), j)$

À la fin de l'algorithme, si j est tel que $l(j)$ est défini, alors $\sigma^{l(j)}$ est un simplexe positif. Donc si à la fin de l'algorithme, C_j est nulle alors σ^j est positif. Donc, si C_j n'est pas nulle, alors $(\sigma^{l(j)}, \sigma^j)$ est une paire de persistence. ■

Définition 2.13 On représente sur un *diagramme de persistence* les *paires de persistence* $\left(\sigma^{l(j)}, \sigma^j\right)$ par $\left(l\left(j\right), j\right)$ ou $\left(f\left(\sigma^{l(j)}\right), f\left(\sigma^j\right)\right)$. On ajoute au diagramme la diagonale $\{y = x\}$ et, pour chaque simplexe positif qui n'est pas dans une paire σ^i, le point $(i, +\infty)$.

Définition 2.14 Si D_1, D_2 sont deux diagrammes (potentiellement augmentés pour avoir le même cardinal) :

La Distance du Goulot est définie par :

$$d_B^\infty(D_1, D_2) = \inf_{\gamma \in \Gamma} \sup_{p \in D_1} \|p - \gamma(p)\|_\infty$$

La Distance p-Wasserstein est définie, pour $p \geq 1$ par :

$$W_p(D_1, D_2) = \inf_{\gamma \in \Gamma} \left(\sum_{\rho \in D_1} \|\rho - \gamma(\rho)\|_p^p \right)^{\frac{1}{p}}$$

Dans les deux cas, Γ est l'ensemble des bijections entre D_1 et D_2.

R Ces deux définitions peuvent être vues comme le coût du transport optimal pour la norme infinie et la norme p. Ces deux définitions sont par ailleurs équivalentes à un facteur près, ce qui est important pour les théorèmes de stabilité.

Théorème 2.15 Si $f, g : X \to \mathbb{R}$ sont *modérées*, on a :

$$d_B^\infty(D_f, D_g) \leq \|f - g\|_\infty$$

où D_φ est le diagramme de persistence de la filtration associée aux ensembles de sous-niveau de φ sur X.

On reviendra plus tard sur la notion de modération.

2.3 Calcul de filtrations

2.3.1 Complexes de Čech, de Vietoris-Rips, et autres

Définition 2.16 On considère un recouvrement $\mathcal{U}$ par des ouverts d'un espace topologique X. Le *complexe de Čech* $C\left(\mathcal{U}\right)$ associé au recouvrement $\mathcal{U}$ vérifie :

- L'ensemble de sommets de $C\left(\mathcal{U}\right)$ est l'ensemble $\mathcal{U}$.

- $[U_0, \dots, U_k]$ est un k-simplexe dans $C(\mathcal{U})$ si et seulement si $\cap U_j \neq \varnothing$.

Théorème 2.17 — Nerveux (Leray) Si toutes les intersections entre les ouverts de $\mathcal{U}$ sont soit vides soit contractibles, alors $C(\mathcal{U})$ et X sont homotopiquement équivalents.

Si on se donne plutôt un nuage de point V dans un espace métrique (X, d) et un réel α.

Définition 2.18 Le *complexe de Čech* $\check{\text{C}}\text{ech}(V, \alpha)$ est le complexe simplicial filtré indexé par $\mathbb{R}$ dont l'ensemblde sommets est V et tel que :

$$\sigma = [p_0, \dots, p_k] \in \check{\text{C}}\text{ech}(V, \alpha) \Leftrightarrow \bigcap_{i=0}^{k} B(p_i, \alpha) \neq \varnothing$$

Définition 2.19 Le *complexe de Vietoris-Rips* $\text{Rips}(V)$ est le complexe simplicial filtré indexé par $\mathbb{R}$ dont l'ensemble de sommets est V et est défini par :

$$\sigma = [p_0, \dots, p_k] \in \check{\text{C}}\text{ech}(V, \alpha) \Leftrightarrow \forall i, j \in \{0, \dots, k\}, d(p_i, p_j) \leq \alpha$$

Proposition 2.20 On a, pour tout $\alpha > 0$:

$$\check{\text{C}}\text{ech}\left(L, \frac{\alpha}{2}\right) \subseteq \text{Rips}(L, \alpha) \subseteq \check{\text{C}}\text{ech}(L, \alpha)$$

Définition 2.21 Si $V = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$, on définit la *cellule de Voronoï* associée à p_i par :

$$\mathcal{V}\wr\nabla(p_i) = \left\{x \in \mathbb{R}^d \,\middle|\, \forall j, \|x - p_i\| \leq \|x - p_j\|\right\}$$

Le *complexe de Delaunay* $\mathcal{D}(P)$ est le nerf de la couverture faite par les cellules de Voronoï. *L'alpha complexe* $\mathcal{A}(P, \alpha)$, pour $\alpha \geq 0$ est le nerf de la famille :

$$(\text{Vor}(p_i) \cap B(p_i, \sqrt{\alpha}))_{i=1,\dots,n}$$

Théorème 2.22 $\mathcal{A}(P, \alpha)$ est homotopie équivalent à $\bigcup_{i=1}^{n} B(p_i, \sqrt{\alpha})$.

2.3.2 Modules de Persistence

On va utiliser ci-dessous la distance de Hausdorff :

$$d_H(A, B) = \max\left\{\sup_{b \in B} d(b, A), \sup_{a \in A} d(a, B)\right\}$$

et la distance de Gromov-Hausdorff :

$$d_{GH}(\mathbb{X}, \mathbb{Y}) = \inf_{\mathbb{Z}, \gamma_1, \gamma_2} (\mathbb{X}, \mathbb{Y})$$

l'infimum étant pris pour $\mathbb{Z}$ un espace métrique, et γ_1, γ_2 des immersions isométriques de $\mathbb{X}, \mathbb{Y}$ dans $\mathbb{Z}$.

Théorème 2.23 Si $\mathbb{X}$ et $\mathbb{Y}$ sont des espaces métriques pré-compacts :

$$d_\infty(\text{Rips}(\mathbb{X}), \text{Rips}(\mathbb{Y})) \leq 2 d_{GH}(\mathbb{X}, \mathbb{Y})$$

Ceci est notamment utile lorsqu'on considère la classification de formes non-rigides, puisqu'alors celles ci sont

presque isométriques, mais que calculer leur distance de Gromov-Hausdorff est très coûteux. On va désormais essayer de démontrer ces résultats de stabilité :

Définition 2.24 Un *module de persistence* $\mathbb{V}$ est une famille d'espaces vectoriels $(V_a)_{a\in\mathbb{R}}$ et une famille $v_a^b : V_a \to V_b, a \leq b$ qui se compose bien et de sorte que v_a^a soit l'identité.

- Si $\mathbb{S}$ est un complexe simplicial filtré, les familles $V_a = H(\mathbb{S}_a)$ et $v_a^b : H(\mathbb{S}_a) \to H(\mathbb{S}_b)$ l'application linéaire induites par l'inclusion $\mathbb{S}_a \hookrightarrow \mathbb{S}_b$ forment un module de persistence.
- Étant donné un espace métrique $\mathbb{X}$, $H(\mathrm{Rips}(\mathbb{X}))$ est un module de persistence.
- La filtration par les sous-niveaux de f induit un module de persistence au niveau de l'homologie.

R Il faut voir un module de persistence comme un foncteur de la (petite) catégorie associée au poset d'indexation, vers la catégorie des modules sur un anneau A. Ici, c'est donc un foncteur de $\mathbb{R}$ vers $\mathrm{Vect}_{\mathbb{F}_2}$, puisqu'on considère notre homologie dans $\mathbb{F}_2$.

Définition 2.25 On dit qu'un module de persistence est dit *q-modéré* si pour tout $a < b$, v_a^b est de rang fini.

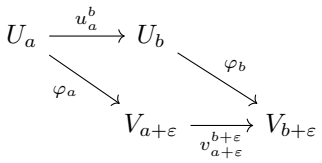
Si $\mathbb{X}$ est pré-compact métrique, alors $H(\mathrm{Rips}(\mathbb{X}))$ et $H(\check{\mathrm{C}}\mathrm{ech}(\mathbb{X}))$ sont q-modérés.

Cette condition apporte de forts théorèmes :

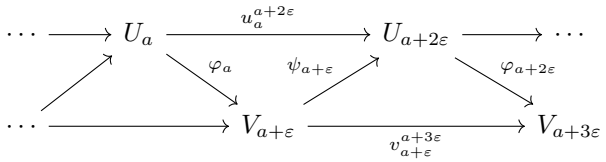
Théorème 2.26 Les modules de persistence q-modérés ont des diagrammes de persistence bien définis.

Il faut ici entendre la notion de diagrammes de persistence comme définis par les bases des espaces.

Définition 2.27 Un *homomorphisme de degré ε* entre deux modules de persistence est une collection Φ d'application linéaire vérifiant :



Un *ε-intercalaire* entre $\mathbb{U}$ et $\mathbb{V}$ est défini par deux homomorphismes de degré ε $\Phi : \mathbb{U} \to \mathbb{V}$ et $\Psi : \mathbb{V} \to \mathbb{U}$ de vérifiant :



Théorème 2.28 Si $\mathbb{U}$ et $\mathbb{V}$ sont q-modérés et ε-entrelacés pour un certain $\varepsilon \geq 0$, alors :

$$d_\infty(\mathrm{diag}(\mathbb{U}), \mathrm{diag}(\mathbb{V})) \leq \varepsilon$$

On va donc chercher à construire des filtrations qui induisent par leurs groupes d'homologie des modules de persistence q-modérés, et qui sont ε-entrelacés quand les espaces/fonctions considérées sont $O(\varepsilon)$-proches. Plus particulièrement, on va démontrer la docilité des complexes de Rips et de Čech.

2.3.3 Théorèmes de Stabilité

Définition 2.29 Une *application multivaluée* C de $\mathbb{X}$ dans $\mathbb{Y}$ est une partie de $\mathbb{X} \times \mathbb{Y}$ qui se projette surjectivement sur $\mathbb{X}$ par la projection $\pi_{\mathbb{X}}$ définissant le produit. *L'image* $C(\sigma)$ de $\sigma \subseteq \mathbb{X}$ est la projection canonique sur $\mathbb{Y}$ de la

préimage de σ par $\pi_{\mathbb{X}}$. La *transposée* ${}^tC$ de C est l'image de C par la symétrie $\mathbb{X} \times \mathbb{Y} \to \mathbb{Y} \times \mathbb{X}$. Si ${}^tC$ est aussi une application multivaluée, on dit que C est une *correspondance.*

Définition 2.30 Si $(\mathbb{X}, \rho_{\mathbb{X}}$ et $(\mathbb{Y}, \rho_{\mathbb{Y}})$ sont des espaces métriques compacts, une correspondance C de $\mathbb{X}$ dans $\mathbb{Y}$ est une *ε-correspondance* si :

$$\forall (x,y),(x',y') \in C, |\rho_X(x,x') - \rho_{\mathbb{Y}}(y,y')| \leq \varepsilon$$

Proposition 2.31 Avec les hypothèses de la définition ci-dessus :

$$d_{GH}(\mathbb{X},\mathbb{Y}) = \frac{1}{2} \inf \{\varepsilon \geq 0 \,|\, \exists C, C \text{ est une } \varepsilon\text{-correspondance}\}$$

Définition 2.32 Si $\mathbb{S}$, $\mathbb{T}$ sont des complexes simpliciaux filtrés avec des ensembles de sommets $\mathbb{X}$ et $\mathbb{Y}$ respectivement, une application multivaluée C de $\mathbb{X}$ dans $\mathbb{Y}$ est dite *ε-simpliciale* de $\mathbb{S}$ dans $\mathbb{T}$ si pour tout $a \in \mathbb{R}$ et tout simplexe $\sigma \in \mathbb{S}_a$, chaque partie finie de $C(\sigma)$ est un simplexe de $\mathbb{T}_{a+\varepsilon}$.

Proposition 2.33 Si C est une correspondance telle que C et ${}^tC$ sont ε-simplicales de $\mathbb{S}$ dans $\mathbb{T}$, elle permette de construire un ε-intercalaire entre $H(\mathbb{S})$ et $H(\mathbb{T})$.

Démonstration. Il suffit pour ça de voir qu'une fonction ε-simpliciale définit immédiatement un homomorphisme de degré ε sur les modules de persistence définis par les filtrations considérées, par continuation linéaire. ■

Proposition 2.34 Si $\mathbb{X}, \mathbb{Y}$ sont des espaces métriques, pour tout $\varepsilon > 2\,\mathrm{d}_{GH}(\mathbb{X},\mathbb{Y})$, alors $H(\mathrm{Rips}(\mathbb{X}))$ et $H(\mathrm{Rips}(\mathbb{Y}))$ sont ε-entrelacés.

Démonstration. Si C est une correspondance de $\mathbb{X}$ dans $\mathbb{Y}$ avec une distortion d'au plus ε. Si $\sigma \in \mathrm{Rips}(\mathbb{X}, a)$ alors $\rho_{\mathbb{X}}(x,x') \leq a$ pour tout $x, x' \in \sigma$. Soit $\tau \subseteq C(\sigma)$ fini. Pour tout $y, y' \in \tau$, il existe $x, x' \in \sigma$ tels que : $y \in C(x)$ et $y' \in C(x')$ donc :

$$\rho_{\mathbb{Y}}(y,y') \leq \rho_{\mathbb{X}}(x,x') + \varepsilon \leq a + \varepsilon$$

Ainsi, $\tau \in \mathrm{Rips}(\mathbb{Y}, a+\varepsilon)$. De même ${}^tC$ est ε-simpliciale de $\mathrm{Rips}(\mathbb{Y})$ dans $\mathrm{Rips}(\mathbb{Y})$. On conclut par la Proposition 2.33. ■

Proposition 2.35 Si $\mathbb{X}, \mathbb{Y}$ sont des espaces métriques, pour tout $\varepsilon \geq 2\,\mathrm{d}_{GH}(\mathbb{X},\mathbb{Y})$, alors $H\left(\check{\mathrm{C}}\mathrm{ech}(\mathbb{X})\right)$ et $H\left(\check{\mathrm{C}}\mathrm{ech}(\mathbb{Y})\right)$ sont ε-entrelacés.

La preuve est similaire à celle d'avant.

Théorème 2.36 Soit $\mathbb{X}$ un espace métrique compact. Les modules de persistence associés à l'homologie des complexes $\mathrm{Rips}(\mathbb{X})$ et $\check{\mathrm{C}}\mathrm{ech}(\mathbb{X})$ sont q-modérés.

Démonstration. On veut montrer que $I_a^b : H(\mathrm{Rips}(\mathbb{X}, a)) \to H(\mathrm{Rips}(\mathbb{X}, b))$ sont de rang finis quand $a < b$. On pose $\varepsilon = (b-a)/2$ et $F \subseteq \mathbb{X}$ un ensemble fini tel que $d_H(\mathbb{X}, F) \leq \varepsilon/2$. Alors :

$$C = \left\{(x,f) \in X \times F \,\middle|\, d(x,f) \leq \frac{\varepsilon}{2}\right\}$$

définit une ε-correspondance. Utilisant l'application d'intercalage de X et F, I_a^b se factorise en :

$$H(\mathrm{Rips}(X,a)) \longrightarrow \underset{\text{de dimension finie}}{H(\mathrm{Rips}(F, a+\varepsilon))} \to H(\mathrm{Rips}(X, a+2\varepsilon)) = H(\mathrm{Rips}(X,b))$$

■

Théorème 2.37 Si $\mathbb{X}$, $\mathbb{Y}$ sont des espaces métriques compacts, alors :

$$d_\infty\left(\mathrm{diag}\left(H\left(\check{\mathrm{C}}\mathrm{ech}\left(\mathbb{X}\right)\right)\right), \mathrm{diag}\left(H\left(\check{\mathrm{C}}\mathrm{ech}\left(Y\right)\right)\right)\right) \leq 2 d_{GH}\left(\mathbb{X}, \mathbb{Y}\right)$$
$$d_\infty\left(\mathrm{diag}\left(H\left(\mathrm{Rips}\left(\mathbb{X}\right)\right)\right), \mathrm{diag}\left(H\left(\mathrm{Rips}\left(Y\right)\right)\right)\right) \leq 2 d_{GH}\left(\mathbb{X}, \mathbb{Y}\right)$$

La preuve des deux derniers théorèmes n'utilise pas l'inégalité triangulaire on pourrait donc étendre les résultats précédents à des espaces munies d'une similarité.

Cependant, on a des problèmes avec la dimension de nos espaces : pour tout $0 < \alpha \leq \beta \in \mathbb{R}$, il existe un espace métrique compact X (immersible dans $\mathbb{R}^4$) tel que pour tout $a \in [\alpha, \beta]$, $H_k(\mathrm{Rips}(X, a))$ est de dimension indénombrable. Toutefois :

- Si X est compact, $\dim H_1(\check{\mathrm{C}}\mathrm{ech}(X, a)) < +\infty$
- Si X est géodésique, $\dim H_1(\mathrm{Rips}(X, a)) < +\infty$ pour $a > 0$ et $\mathrm{diag}(H_1(\mathrm{Rips}(X)))$ est contenu dans la ligne $x = 0$
- Si X est un espace géodésique δ-hyperbolique, $\mathrm{diag}(H_2(\mathrm{Rips}(X)))$ est contenu dans une bande verticale de largeur $\mathcal{O}(\delta)$.

3 Homologie pratique, échantillonnage et apprentissage

3.1 Résilience statistique

Le complexe de Vietoris-Rips et ses filtrations se calculent en $\mathcal{O}\left(|\mathbb{X}|^d\right)$, ce qui rend le calcul de persistence quasi impossible en pratique. Par ailleurs, les filtrations et la distance de Gromov-Hausdorff sont très sensibles au bruit et aux anomalies.

3.1.1 Calcul statistique

On va s'intéresser à un espace métrique $(\mathbb{M}, \rho)$ et à une mesure de probabilité μ à support compact X_μ dans $\mathbb{M}$. On échantillonne m points selon μ, ce qui nous donne un nuage de point $\hat{\mathbb{X}}_m$, et une filtration $\mathrm{Filt}\,\hat{\mathbb{X}}_m$. On a alors :

Proposition 3.1 Si $\varepsilon > 0$:

$$\mathbb{P}\left(d_\infty\left(\mathrm{diag}\left(\mathrm{Filt}\left(\mathbb{X}_\mu\right)\right), \mathrm{diag}\left(\mathrm{Filt}\left(\hat{\mathbb{X}}_m\right)\right)\right) > \varepsilon\right) \leq \mathbb{P}\left(d_{GH}\left(\mathbb{X}_\mu, \hat{\mathbb{X}}_m\right) > \frac{\varepsilon}{2}\right)$$

Démonstration. Conséquence directe du Théorème 2.28 de stabilité. ■

On obtient quasi immédiatement des inégalités de déviation :

Définition 3.2 Pour $a, b > 0$, on dit que μ vérifie *la supposition (a, b)-standard* si pour $x \in \mathbb{X}_\mu$ et $r > 0$, on a :

$$\mu(B(x, r)) \geq \min(a r^b, 1)$$

On note $\mathcal{P}(a, b, \mathbb{M})$ *l'ensemble des distributions de probabilité (a, b)-standard* sur $\mathcal{M}$.

Théorème 3.3 Si μ vérifie la supposition (a, b)-standard, pour tout $\varepsilon > 0$:

$$\mathbb{P}\left(d_\infty\left(\mathrm{diag}\left(\mathrm{Filt}\left(\mathbb{X}_\mu\right)\right), \mathrm{diag}\left(\mathrm{Filt}\left(\hat{\mathbb{X}}_m\right)\right)\right) > \varepsilon\right) \leq \min\left(\frac{8^b}{a\varepsilon^b} \exp\left(-ma\varepsilon^b\right), 1\right)$$

De plus :

$$\mathbb{P}\left(d_\infty\left(\mathrm{diag}\left(\mathrm{Filt}\left(\mathbb{X}_\mu\right)\right), \mathrm{diag}\left(\mathrm{Filt}\left(\hat{\mathbb{X}}_m\right)\right)\right) \geq C_1\left(\frac{\log m}{m}\right)^{1/b}\right) \xrightarrow[m\to\infty]{} 1$$

où C_1 est une constante qui ne dépend que de a et b.

Démonstration. On commence par majorer $\mathbb{P}\left(\mathrm{d}_{GH}\left(\mathbb{X}_\mu, \hat{\mathbb{X}}_m\right) > \frac{\varepsilon}{2}\right)$ puis, on obtient par la supposition (a,b)-standard une borne supérieure explicite pour la couverture de $\mathbb{X}_\mu$ par des boules de rayon $\varepsilon/2$. On peut alors conclure en prenant l'union des bornes. ■

Théorème 3.4 On a :

$$\sup_{\mu\in\mathcal{P}(a,b,\mathbb{M})} \mathbb{E}[d_\infty(\mathrm{diag}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathrm{diag}(\mathrm{Filt}(\hat{\mathbb{X}}_m)))] \leq C(\frac{\ln m}{m})^{1/b}$$

où C ne dépend que de a et b. Si de plus il y a un point non isolé x dans $\mathbb{M}$, et si $x_m \in \mathbb{M}\setminus\{x\}$, telle que $\rho(x,x_m) \leq (am)^{-1/b}$, pour tout estimateur $\hat{\mathrm{diag}}_m$ de $\mathrm{diag}(\mathrm{Filt}(\mathbb{X}_\mu))$:

$$\liminf_{m\to\infty} \rho(x,x_m)^{-1} \sup_{\mu\in\mathcal{P}(a,b,\mathbb{M})} \mathbb{E}[d_\infty(\mathrm{diag}(\mathrm{Filt}(\mathbb{X}_\mu)), \hat{\mathrm{diag}}_m)] \geq C'$$

où C' est une constante absolue.

3.1.2 Paysages de persistence et rééchantillonnage

Définition 3.5 Si on a un diagramme de persistence (b_i, d_i), son *paysage de persistence* est obtenu en y ajoutant à chaque point les deux projections orthogonales sur la diagonale par rapport aux axes, puis en plaçant à l'horizontale sa diagonale. Formellement, c'est l'union pour $p = (\frac{b+d}{2}, \frac{d-b}{2})$ des graphes :

$$\Lambda_p(t) = \begin{cases} t-b & t \in [b, \frac{b+d}{2}] \\ d-t & t \in [\frac{b+d}{2}, d] \\ 0 & \text{sinon} \end{cases}$$

C'est un encodage de la persistence comme un élément d'un espace fonctionnel, comme fait ci-dessous :

Définition 3.6 On définit le *k-ème paysage* d'un diagramme D par :

$$\lambda_D(k,t) = \underset{p\in D}{\mathrm{kmax}} \Lambda_p(t), t \in \mathbb{R}, k \in \mathbb{N}$$

où kmax désigne le k-ème plus grand élément d'un ensemble.

Proposition 3.7

- Pour $t \in \mathbb{R}$ et $k \in \mathbb{N}$, $0 \leq \lambda_D(k,t) \leq \lambda_D(k+1,t)$
- Pour $t \in \mathbb{R}$, et $k \in \mathbb{N}$, $|\lambda_D(k,t) - \lambda_{D'}(k,t)| \leq d_\infty(D,D')$

Dans la suite, on note $\mathcal{L}_T$ les paysages dont le support est dans $[0,T]$, on prend P une distribution de probabilité sur $\mathcal{L}_T$ et $\lambda_1, \ldots, \lambda_n \sim P$ i.i.d. On note $\mu(t) = \mathbb{E}[\lambda_i(t)]$ le paysage moyen et on l'estime par la moyenne échantillonnée :

$$\bar{\lambda}_n(t) = \frac{1}{n}\sum \lambda_i(t)$$

$\bar{\lambda}_n$ est un estimateur point à point non biaisé de μ, qui converge point à point.

Définition 3.8 Soit $\mathcal{F}$ la famille des applications d'évaluation $f_t : \mathcal{L}_T \to \mathbb{R}$. Le *processus empirique* indexé par les f_t est défini par :

$$\mathbb{G}_n(t) = \sqrt{n}(\bar{\lambda}_n(t) - \mu_t) = \sqrt{n}(P_n - P)(f_t)$$

Théorème 3.9 Soit $\mathbb{G}$ un pont Brownien avec fonction de covariance :

$$\kappa(s,t) = \int f_t(\lambda) f_s(\lambda)\,\mathrm{d}P(\lambda) - \int f_t(\lambda)\,\mathrm{d}P(\lambda) \int f_s(\lambda)\,\mathrm{d}P(\lambda).$$

On a alors :

$$\mathbb{G}_n \to \mathbb{G}$$

pour la convergence faible.

Si de plus on note $\sigma(t)$ l'écart-type de $\sqrt{n}\bar{\lambda}_n(t)$:

Théorème 3.10 Si $\sigma(t) > c > 0$ sur un intervalle $I = [t_*, t^*] \subseteq [0,T]$ pour une constante c, avec $W = \sup_{t\in I} |\mathbb{G}(f_t)|$ on a :

$$\sup_{z\in\mathbb{R}} \left| \mathbb{P}\left(\sup_{t\in[t_*,t^*]} |\mathbb{G}_n(t)| \le z \right) - \mathbb{P}(W \le z) \right| = \mathcal{O}\left(\frac{(\log n)^{7/8}}{n^{1/8}} \right)$$

C'est une forme de théorème central limite uniforme. On a de plus le corollaire suivant :

Théorème 3.11 Sous les mêmes hypothèses, étant donné un niveau de confiance $1-\alpha$, on peut construire des fonctions de confiance $l_n(t)$ et $u_n(t)$ telles que :

$$\mathbb{P}(l_n(t) \le \mu(t) \le u_n(t), \forall t \in I) \ge 1 - \alpha - \mathcal{O}\left(\frac{(\log n)^{7/8}}{n^{1/8}} \right)$$

De plus, on a :

$$\sup_t u_n(t) - l_n(t) = \mathcal{O}\left(\sqrt{\frac{1}{n}} \right)$$

Autrement dit, le rééchantillonnage (ou *bootstrap*) permet d'obtenir des intervalles de confiance pour les paysages.

3.1.3 Bruit et méthode d'échantillonnage

On va ici s'intéresser à l'impact de la procédure d'échantillonnage. On rappelle que la définition des distances de p-Wasserstein peuvent se poser pour n'importe quelles deux mesures de probabilité sur un même espace métrique $(\mathbb{M}, \rho)$. On a notamment le théorème suivant :

Théorème 3.12 Si μ, ν sont des mesures de probabilité sur un même espace métrique $(\mathbb{M}, \rho)$, on a :

$$\|\Lambda_{\mu,m} - \Lambda_{\nu,m}\|_\infty \le m^{\frac{1}{p}} \mathrm{W}_p(\mu,\nu)$$

où W_p dénote la distance de Wasserstein associée à la fonction de coût $\rho(\cdot,\cdot)^p$.

Ceci nous assure de l'utilisabilité des méthodes d'échantillonnage, et notamment de la robustesse aux échantillons peu probables et des méthodes sous-échantillonnant.

Avant de démontrer ceci, donnons trois courts lemmes sur les passages des espaces d'échantillonnage aux espaces de paysages :

Lemme 3.13 Pour toutes mesures de probabilité μ, ν sur $(\mathbb{M}, \rho)$, si ρ_m est une métrique sur $\mathbb{M}^m$ telle que :

$$\rho_m(X,Y) \leq \left(\sum_{i=1}^{m} \rho(x_i, y_i)^p\right)^{\frac{1}{p}}$$

alors :

$$\mathrm{W}_p(\mu^{\otimes m}, \nu^{\otimes m}) \leq m^{\frac{1}{p}}\, \mathrm{W}_p(\mu, \nu)$$

Démonstration. Si Π est un plan de transport entre μ et ν, alors $\Pi^{\otimes m}$ est un plan de transport entre $\mu^{\otimes m}$ et $\nu^{\otimes m}$ et donc :

$$\begin{aligned}
&\int_{\mathbb{M}^{2m}} \rho_m(X,Y)^p \,\mathrm{d}\Pi^{\otimes m}(X,Y) \leq \int_{\mathbb{M}^m \times \mathbb{M}^m} \sum_{i=1}^{m} \rho(x_i, y_i)^p \,\mathrm{d}\Pi(x_1, y_1) \cdots \mathrm{d}\Pi(x_m, y_m) \\
&= m \int_{\mathbb{M}\times\mathbb{M}} \rho(x_1, y_1)^p \,\mathrm{d}\Pi(x_1, y_1),
\end{aligned}$$

ce qui conclut la preuve. ■

Lemme 3.14 En notant $\varphi^m : \mathbb{M}^m \to \mathcal{D}$ la fonction qui à X associe $\mathrm{diag}\,(\mathrm{Filt}\, X)$ dans l'espace des diagrammes de persistence et si Φ_μ^m est le poussé en avant de μ par φ^m :

$$\mathrm{W}_p(\Phi_\mu^m, \Phi_\nu^m) \leq \mathrm{W}_p(\mu^{\otimes m}, \nu^{\otimes m})$$

Démonstration. En notant $\Delta_m(X,Y) = (\psi(\varphi^m(X)), \psi(\varphi^m(Y)))$, si Π est un plan entre $\mu^{\otimes m}$ et $\nu^{\otimes m}$, alors le plan $\Delta_m \sharp \Pi$ poussé en avant de Π est un plan entre Φ_μ^m et Φ_ν^m et on a :

$$\begin{aligned}
\int_{\mathcal{D}^2} \mathrm{W}_\infty(D_X, D_Y)^p \,\mathrm{d}\Delta_m\sharp\Pi(D_X, D_Y) &= \int_{\mathbb{M}^{2m}} \mathrm{W}_\infty(\varphi^m(X), \varphi^m(Y))^p \,\mathrm{d}\Pi(X,Y) \\
&\leq \int_{\mathbb{M}^{2m}} d_H(X,Y)^p \,\mathrm{d}\Pi(X,Y) \quad (2.23) \\
&\leq \int_{\mathbb{M}^{2m}} \rho_m(X,Y)^p \,\mathrm{d}\Pi(X,Y)
\end{aligned}$$

■

Lemme 3.15 En notant ψ l'application de l'espace des diagrammes vers l'espace des paysages munis de la norme infinie et Ψ_μ^m le poussé en avant de φ_μ^m par ψ :

$$\left\| \mathbb{E}_{\lambda_X \sim \Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\lambda_Y \sim \Psi_\nu^m}[\lambda_Y] \right\|_\infty \leq \mathrm{W}_{\infty,p}(\Phi_\mu^m, \Phi_\nu^m)$$

où $\mathrm{W}_{\infty,p}$ fait appel à la p-ème puissance de la distance infinie de Wasserstein pour les diagrammes sous-jacents.

Démonstration. Si Π est un plan entre Φ_μ^m et Φ_ν^m, pour tout $t \in \mathbb{R}$ on a :

$$\begin{aligned}
\left| \mathbb{E}_{\lambda_X \sim \Psi_\mu^m}[\lambda_X](t) - \mathbb{E}_{\lambda_Y \sim \Psi_\nu^m}[\lambda_Y](t) \right|^p &= \left| \mathbb{E}\left[\lambda_X(t) - \lambda_Y(t)\right] \right|^p \\
&\leq \mathbb{E}\left[|\lambda_X(t) - \lambda_Y(t)|^p \right] \quad \text{(Jensen)} \\
&\leq \mathbb{E}\left[\mathrm{W}_\infty(D_X, D_Y)^p \right] \quad (3.7) \\
&= \int_{\mathcal{D}\times\mathcal{D}} \mathrm{W}_\infty(D_X, D_Y)^p \,\mathrm{d}\Pi(D_X, D_Y)
\end{aligned}$$

■

Preuve du Théorème 3.12. Par le Lemme 3.13 ci-dessus

$$\mathrm{W}_p\left(\mu^{\otimes m}, \nu^{\otimes m}\right) \leq m^{\frac{1}{p}}\, \mathrm{W}_p(\mu, \nu)$$

Par le Lemme 3.14, en notant P_π le diagramme de persistence associé à π :

$$\mathrm{W}_p(P_\mu, P_\nu) \leq \mathrm{W}_p(\mu^{\otimes m}, \nu^{\otimes m})$$

Enfin, par le Lemme 3.15 :

$$\|\Lambda_{\mu,m} - \Lambda_{\nu,m}\|_\infty \leq \mathrm{W}_p(P_\mu, P_\nu)$$

■

3.2 Persistence et apprentissage automatique

3.2.1 Représentation de persistence

Puisque l'espace des diagrammes de persistence n'est pas linéaire, les algorithmes de ML classique ne fonctionnent pas bien. La bibliothèque Python et C++ *Gudhi* propose une large zoologie de représentations pour la persistence, comme mesures discrètes, espaces métriques finis, racines de polynômes ou collections de fonctions 1D.

Par exemple, on peut représenter un diagramme en le plongeant dans $\mathbb{R}^2$ et l'espace des mesures par $D = \sum \delta_{p_i}$.

Définition 3.16 Si on se donne un noyau $K : \mathbb{R}^2 \to \mathbb{R}$ et H une matrice de bande-passante (forme quadratique), en définissant $K_H(u) = |H|^{-1/2} K(H^{-1/2} \cdot u)$, on obtient alors, étant donné une fonction de poids w, *la surface de persistence* de D par :

$$\forall u \in \mathbb{R}^2, \rho(D)(u) = D(wK_H(u - \cdot))$$

La question se pose alors de savoir comment choisir une représentation adaptée à un réseau de neurones. Une réponse partielle peut être trouvée en regardant l'architecture à ensembles profonds : on se donne n points dans $\mathbb{R}^d$ et on construit un réseau dont les niveaux sont invariants par permutation ($f \circ \sigma = f$)

Théorème 3.17 — Universalité Une fonction f est invariante par permutation si et seulement si, pour tout X inclus dans un ensemble dénombrable $f(X) = \rho(\sum_i \varphi(x_i))$ pour certaines fonctions ρ et φ.

Les réseaux à niveaux invariants par permutation permettent de généraliser plusieurs approches générales en TDA, sous la forme de "niveaux de persistence"

$$\mathrm{PersLay}(\mathrm{diag}) = \rho(\mathrm{op}\{w(p), \varphi(p)\}_{p \in \mathrm{diag}}),$$

où op est invariante par permutation, w est une fonction de poids, et φ est une transformation permettant de se ramener à un ensemble dénombrable.

On peut par exemple retrouver la surface de persistence en se donnant $t_1, \ldots, t_q \in \mathbb{R}^2$ puis en posant :

- $w(p) = w_t((x, y))$;
- $\varphi_\Gamma : p \mapsto (\Gamma_p(t_i))_i$ avec Γ_p la gaussienne centrée en p d'écart-type fixé σ ;
- $\mathrm{op} = \sum$.

Pour les paysages, on prend $w(p) = 1$, $\mathrm{op} = \mathrm{top}{-}k$ et φ_Λ l'évaluation de Λ_p en q paramètres $t_1, \ldots, t_q$.

3.2.2 Différentiabilité de la persistence

Nombre de méthodes permettent de minimiser une fonction sur l'ensemble des diagrammes, mais la plupart sont restreintes à un type spécifique de filtration ou de fonction à minimiser.

Définition 3.18 Étant donné un ensemble V, un complexe simplicial K sur V et une filtration K_r indexée par un ensemble $R \subseteq \mathbb{R}$, pour $\sigma \in K$ on pose

$$\Phi_\sigma = \inf \{r \in R \,|\, \sigma \in K_r\}.$$

Par conséquent, une filtration de K est un vecteur $|K|$-dimensionnel $\Phi = (\Phi_\sigma)_\sigma \in \mathbb{R}^{|K|}$ tel que $\tau \subseteq \sigma \Rightarrow \Phi_\tau \leq \Phi_\sigma$.

Proposition 3.19 L'ensemble $\mathrm{Filt}_K \subseteq \mathbb{R}^{|K|}$ des vecteurs sur $\mathbb{R}^{|K|}$ définissant une filtration sur K est semi-algébrique sur $\mathbb{R}$.

Définition 3.20 Si K est un complexe simplicial, une application $\Phi : A \to \mathbb{R}^{|K|}$ est une *famille paramétrée de filtrations* si pour tout $\tau \subseteq \sigma$, on a $\Phi_\tau \leq \Phi_\sigma$.

Le calcul de l'homologie persistente dans ce cas se fait avec l'Algorithme 1, et peut donc se voir comme suit :

Filtration — Appairage de Simplexes — Diagramme de Persistence

$$\Phi = (\Phi_\sigma)_{\sigma \in K} \in \mathbb{R}^{|K|} \longrightarrow p \text{ paires } (\sigma_{l(i)}, \sigma_i),\ q\ \sigma_l \text{ esseulés avec } |K| = 2p + q \longrightarrow (\Phi_{\sigma_{l(i)}}, \Phi_{\sigma_i}), (\Phi_{\sigma_l}, \infty)$$

$$\downarrow \text{ ordre lexicographique}$$

$$D(\Phi) \in \mathbb{R}^{|K|}$$

$$\mathrm{Pers}$$

L'application de persistence Pers correspond à une permutation des coordonnées localement constante. On va chercher à utiliser cette application pour restreindre l'ensemble des filtrations afin de pouvoir différentier l'application Pers. Pour cela, on rappelle la notion de structure o-minimale :

Définition 3.21 Une *structure o-minimale* sur le corps des réels $\mathbb{R}$ est une collection $(S_n)_{n \in \mathbb{N}}$ où chaque $S_n \in \mathcal{P}(\mathcal{P}(\mathbb{R}))$ est un ensemble de parties de $\mathbb{R}^n$ tel que

1. S_1 est exactement la collection des unions finies de points et d'intervalles ;
2. Les parties algébriques de $\mathbb{R}^n$ sont dans S_n ;
3. S_n est une sous-algèbre booléenne de $\mathbb{R}^n$ pour tout $n \in \mathbb{N}$;
4. Si $A \in S_n$ et $B \in S_m$ alors $A \times B \in S_{n+m}$;
5. Si $\pi^n_{n+1} : \mathbb{R}^{n+1} \to \mathbb{R}^n$ est la projection sur les n premières coordonnées, alors $\pi(S_{n+1}) \subseteq S_n$;

$A \in S_n$ est dit *définissable* dans la structure o-minimale. Pour $A \subseteq \mathbb{R}^n$, $f : A \to \mathbb{R}^m$ est une application définissable si son graphe est définissable dans $\mathbb{R}^{n+m}$.

Définition 3.22 Une *stratification* d'un espace topologique est une filtration finie par ensembles fermés F_i telles que la différence entre deux membres successifs de la filtration F_i et F_{i-1} est soit vide soit une sous-variété lisse de dimension i. Les composantes connexes de la différence $F_i \setminus F_{i-1}$ sont les *strates* de dimension i. Une stratification (dans $\mathbb{R}^n$) vérifie les *propriétés de Whitney* si toute paire de strates vérifient les deux conditions suivantes :

1. X et Y vérifient la condition A de Whitney si, lorsqu'une suite x_m de X converge vers $y \in Y$ et lorsque la suite des i-plans tangents T_m à X en x_m converge vers un i-plan T, alors T contient le j-plan tangent à Y en y ;
2. X et Y vérifient la condition B de Whitney si, pour toute suite x_m de X et toute suite y_m de Y qui convergent vers le même point y de Y de sorte que la suite de ligne sécantes L_m entre x_m et y_m converge vers une ligne L et que la suite de i-plans tangents T_m à X en x_m converge vers un i-plan T alors L est contenue dans T.

Proposition 3.23 Tous les ensembles définissables admettent des stratifications finies vérifiant les propriétés de Whitney.

Proposition 3.24 Pour un complexe simplicial K, l'application

$$\mathrm{Pers} : \mathrm{Filt}_K \subseteq \mathbb{R}^{|K|} \to \mathbb{R}^{|K|}$$

est semie-algébrique (et donc définissable dans toute structure o-minimale). De plus, il existe une partition semie-algébrique de Filt_K telle que la restriction de Pers à chaque élément de la partition est Lipschitz.

Corollaire 3.25 Si K est un complexe simplicial et $\Phi : A \to \mathbb{R}^{|K|}$ est une famille paramétrée de filtrations définissable dans une structure o-minimale donnée, alors $\mathrm{Pers} \circ \Phi$ est définissable.

Il faut entendre définissable au sens de définissable dans une structure o-minimale donnée. Puisque les ensembles semis algébriques définissent une structure o-minimale on peut toujours remplacer définissable par semi-algébrique.

Proposition 3.26 Si K est un complexe simplicial et Φ est une famille paramétrée définissable sur A de dimension finie m, il existe une partition finie définissable de A notée $S \sqcup O_1 \sqcup \cdots \sqcup O_k$ telle que $\dim S < \dim A = m$ et pour tout $i \leq k$, O_i est une variété définissable de dimension m sur laquelle $(\mathrm{Pers} \circ \Phi)_{|O_i} : O_i \to \mathbb{R}^{|K|}$ est différentiable.

Dans le cas de la filtration de Vietoris-Rips, on prend $\Phi : A = (\mathbb{R}^d)^n \to \mathbb{R}^{|\Delta_n|}$ pour Δ_n le complexe simplicial des faces sur simplexe $(n-1)$-dimensionnel et pour $x = (x_1, \ldots, x_n) \in A$ et tout simplexe $\sigma \subseteq [\![1, n]\!]$, on a

$$\Phi_\sigma(x) = \max_{i,j \in \sigma} \|x_i - x_j\|$$

Pour K un complexe simplicial avec n sommets $v_1, \ldots, v_n$, on pose $\Phi : \mathbb{R}^n \to \mathbb{R}^K$ et pour toute fonction f et tout simplexe σ

$$\Phi_\sigma(f) = \max_{i \in \sigma} f(v_i)$$

3.2.3 Fonctions sur la persistence

Définition 3.27 Une fonction

$$E : \mathbb{R}^{|K|} = \left(\mathbb{R}^2\right)^p \times \mathbb{R}^q \to \mathbb{R}$$

est une *fonction de persistence* si elle est invariante aux permutations des points du diagramme de persistence : pour toutes permutations $\sigma, \sigma' \in \mathfrak{S}_p \times \mathfrak{S}_q$, $E \circ (\sigma \otimes \sigma') = E$ (pour $\otimes$ le produit tensoriel dans Set).

Proposition 3.28 Soit E une fonction de persistence.

- Si E est localement Lipschitz, alors $E \circ \mathrm{Pers}$ est localement Lipschitz.
- Si E et $\Phi : A \subseteq \mathbb{R}^d \to \mathbb{R}^{|K|}$ sont définissables, alors $\mathcal{L} = E \circ \mathrm{Pers} \circ \Phi : A \to \mathbb{R}$ a une sous-différentielle de Clarke

$$\partial \mathcal{L}(z) = \mathrm{Conv}\left\{ \lim_{z_i \to z} \nabla \mathcal{L}(z_i) \,\middle|\, \mathcal{L} \text{ est différentiable en } z_i \right\}$$

bien définie.

Par exemple, l'application de persistence totale $E(D) = \sum_{i=1}^{p} |d_i - b_i|$ est Lipschitz et semie-algébrique. De même, la distance du goulot (ou ∞-Wasserstein) $E(D) = d_B(D, D^*) = \min_m \max_{(p,p^*)\in m} \|p - p^*\|_\infty$, le minimum étant pris sur les couplages partiels entre D et D^*, est semie-algébrique et Lipschitz.

L'existence d'une sous-différentielle/d'un sous-gradient nous permet d'utiliser l'algorithme itératif classique de descente stochastique de sous-gradient

$$x_{k+1} = x_k - \alpha_k \left(y_k + \zeta_k\right), y_k \in \partial\mathcal{L}(x_k), \tag{PersSGD}$$

où la suite (α_k) est le taux d'apprentissage et (ζ_k) est une suite de variables aléatoires de bruit.

On va supposer les hypothèses classiques suivantes :

1. $\alpha_k \geq 0, \sum \alpha_k = +\infty, \sum \alpha_k^2 < +\infty$;
2. $\sum \|x_k\| < \infty$ presque sûrement ;
3. Si $\mathcal{F}_k$ est la suite croissante de σ-algèbres engendrées par les x_j, y_j, ζ_j pour $j < k$, il existe une fonction $p : \mathbb{R}^d \to \mathbb{R}$ bornée sur les ensembles bornés telle que presque sûrement

$$\mathbb{E}\left[\zeta_k \,|\, \mathcal{F}_k\right] = 0 \quad \mathbb{E}\left[\|\zeta_k\|^2 \,\middle|\, \mathcal{F}_k\right] < p(x_k).$$

On obtient alors le théorème de convergence suivant :

Théorème 3.29 Soit K un complexe simplicial, $A \subseteq \mathbb{R}^d$ et $\Phi : A \to \mathbb{R}^{|K|}$ une famille paramétrée de filtrations de K définissable dans une structure o-minimale. Si E est une fonction de persistence définissable telle que $\mathcal{L} = E \circ \mathrm{Pers} \circ \Phi$ est localement Lipschitz, alors, sous les hypothèses ci-dessus, presque sûrement, les points limites de la suite (x_k) obtenus par les itérations de l'Équation (PersSGD) sont des points critiques de $\mathcal{L}$ et la suite $(\mathcal{L}(x_k))$ converge.

3.2.4 Densité de diagrammes de persistence espérés

On s'intéresse ici à une nuage de points V pondéré par $w : V \to \mathbb{R}$.

Définition 3.30 Le *complexe pondéré de Vietoris-Rips* noté $\mathrm{Rips}_w(V)$ est le complexe simplicial filtré indicé par $\mathbb{R}$ dont l'ensemble de sommet est V et défini par

$$\sigma \in [p_0 \cdots p_k] \in \mathrm{Rips}_w(V, \alpha) \Leftrightarrow d(p_i, p_j) \leq \alpha \wedge w(p_i) \leq \alpha$$

On peut donc s'intéresser au calcul de diagrammes espérés, et à la densité (en tant que mesure de probabilité) que l'espérance de diagrammes par rapport à des familles pondérées : si $\mathbb{X}$ est une variable aléatoire sur M^n, et si $\mathcal{K}$ est une fonction de filtration, que dire de $\mathbb{E}\left[\mathcal{D}\left[\mathcal{K}\left(\mathbb{X}\right)\right]\right]$ dans le cas non-asymptotique ($|\mathbb{X}| = n$ fixé/borné) ?

Pour ceci, on va simplement modifier la définition d'une filtration dans ce cas, comme dans la définition 3.18 :

Définition 3.31 Soit $n > 0$ entier, $\mathcal{F}_n$ l'ensemble des parties non-vides de $[\![1, n]\!]$ et M une d-variété connexe lisse possiblement avec bord. Une fonction de filtration

$$\varphi = (\varphi[J])_{J\in\mathcal{F}_n} : M^n \to \mathbb{R}^{|\mathcal{F}_n|}$$

est une fonction qui est

- invariante par permutation : si $\tau \in \mathfrak{S}_n$ a support dans J,

$$\varphi[J](x_{\tau(i)}) = \varphi[J](x_i);$$

- monotone : $J \subseteq J' \Rightarrow \varphi[J] \leq \varphi[J']$.

Sur $x = (x_1, x_n)$, $\varphi(x)$ induit un ordre sur les faces du $(n-1)$-simplexe, c'est à dire une filtration $\mathcal{K}(x)$:

$$\forall J \in \mathcal{F}_n, J \in \mathcal{K}(x, r) \Leftrightarrow \varphi[J](x) \leq r$$

Dans la suite on note, pour $J \in \mathcal{F}_n$ un simplexe et x un vecteur de M^n, $x(J) = (x_j)_{j \in J}$.

En général, on a 5 hypothèses de base :

(K1) Absence d'interaction : si $J \subseteq \mathcal{F}_n$, $\varphi[J](x)$ ne dépend que de $x(J)$;

(K2) Invariance par permutation ;

(K3) Monotonie ;

(K4) Compatibilité : si $j \in J$ un simplexe de $\mathcal{F}_n$, si $\varphi[J](x_1, \ldots, x_n)$ n'est pas fonction de x_j sur un ouvert de U, alors $\varphi[J] \equiv \varphi[J \setminus \{j\}]$ sur U ;

(K5) Régularité : φ est sous-analytique et le gradient de toutes ses entrées est non-nul presque sûrement.

On a également une hypothèse (K5') de régularité, quand le gradient de φ sur chaque entrée J de taille strictement supérieure à 1 est non-nul presque sûrement et que $\varphi[\{j\}] = 0$ pour tout j.

On vérifie facilement que la filtration de Vietoris-Rips vérifie les hypothèses (avec la régularité affaiblie), par exemple.

Définition 3.32 Soit $k \geq 0$. Pour $A \subseteq \mathbb{R}^d$ et $\delta > 0$, on considère

$$\mathcal{H}_k^\delta(A) = \inf \left\{ \sum_i \operatorname{diam}(U_i)^k \,\middle|\, A \subseteq \bigcup_i U_i \wedge \operatorname{diam}(U_i) < \delta \right\}.$$

La *mesure de Hausdorff k-dimensionnelle* sur $\mathbb{R}^D$ de A est définie par

$$\mathcal{H}_k(A) = \lim_{\delta \to 0} \mathcal{H}_k^\delta(A).$$

Théorème 3.33 Soit $n \geq 1$. On suppose que M est une d-variété connexe réelle lisse compacte potentiellement avec bord, que $\mathbb{X}$ est une variable sur M^n avec densité par rapport à la mesure de Hausdorff $\mathcal{H}_{dn}$, et que $\mathcal{K}$ vérifie les hypothèses (K1) à (K5) ci-dessus. Alors, pour tout $s \geq 0$, $\mathbb{E}[\mathcal{D}_s[\mathcal{K}(\mathbb{X})]]$ a une densité par rapport à la mesure de Lebesgue sur le demi-plan $x \leq y$.

Si $\mathcal{K}$ vérifie les hypothèses (K1)-(K4) et (K5'), alors pour tout $s \geq A$, $\mathbb{E}[\mathcal{D}_s[\mathcal{K}(\mathbb{X})]]$ a une densité par rapport à la mesure de Lebesgue sur le demi-plan $x \leq y$, et de plus, $\mathbb{E}[\mathcal{D}_0[\mathcal{K}(\mathbb{X})]]$ a densité par rapport à la mesure de Lebesgue sur la ligne verticale $\{0\} \times [0, \infty)$.

Théorème 3.34 Sous les hypothèses du théorème précédent, si $\mathbb{X}$ a densité de classe $\mathcal{C}^k$ par rapport à $\mathcal{H}_{nd}$, alors de plus pour $s \geq 0$, la densité de $\mathbb{E}[\mathcal{D}_s[\mathcal{K}(\mathbb{X})]]$ est $\mathcal{C}^k$.

Avant de prouver ces théorèmes (ou du moins de donner une esquisse de la preuve), rappelons quelques notion de sous-analyticité.

Définition 3.35 Soit $M \subseteq \mathbb{R}^d$ une d-sous-variété connexe réelle lisse possiblement avec bord.

- $X \subseteq M$ est *semi-analytique* si tout $p \in M$ a un voisinage U_p tel que

$$X \cap U_p = \bigcup_{i=1}^{p} \bigcap_{j=1}^{q} X_{i,j},$$

où les $X_{i,j}$ s'écrivent $f_{i,j}^{-1}(\{0\})$ ou $f_{i,j}^{-1}((0,\infty))$ où les $f_{i,j}$ sont analytiques.

- $X \subseteq M$ est *sous-analytique* si tout point de M admet un voisinage U, une variété lisse N et un ensemble semi-analytique A relativement compact de $N \times M$ tel que $X \cap U$ est la projection de A sur M.
- $f : X \to \mathbb{R}$ est *sous-analytique* si son graphe est sous-analytique sur $M \times \mathbb{R}$. On note $\mathcal{S}(X)$ l'ensemble des fonctions réelles sous-analytiques sur X.
- $x \in X \subseteq M$ est *lisse de dimension k* si, dans un certain voisinage de x dans M,X est une k-sous-variété analytique.
- La *dimension* de X est la dimension maximale d'un point lisse de X.
- On note $\mathrm{Reg}(X)$ les points réguliers de X, i.e. les points lisses de X de dimension d. C'est une partie ouverte de M, potentiellement vide. On note $\mathrm{Sing}(X)$ les points singuliers de X.

Ceci nous permet de dériver quelques lemmes (sans démonstration) :

Lemme 3.36 Pour $f \in \mathcal{S}(M)$, l'ensemble $A(f)$ sur lequel f est analytique est une partie ouverte sous-analytique de M. Son complémentaire est sous-analytique de dimension $< d$.

Lemme 3.37 Si $f, g : X \to \mathbb{R}$ sont sous-analytiques sur une partie sous-analytique de M telles que l'image d'un borné est bornée, alors fg et $f+g$ sont sous-analytiques, et les parties $f^{-1}(\{0\})$ et $f^{-1}((0,\infty))$ sont sous-analytiques dans M.

Lemme 3.38 Soit X sous-analytique dans M. Si la dimension de X est strictement inférieure à d, alors $\mathcal{H}_d = 0$

Corollaire 3.39 On a :

- $\mathcal{H}_d(X) = \mathcal{H}_d(\mathrm{Reg}(X))$;
- Pour toute $f \in \mathcal{S}(M)$, son gradient est défini partout sauf sur un ensemble sous-analytique de dimension strictement inférieure à d (et donc de mesure de Hausdorff nulle).

On va de plus avoir besoin de la formule suivante pour définir la densité de nos diagrammes :

Théorème 3.40 — Formule de la Co-Aire Soit M (respectivement N) une m-variété riemanienne de dimension m (respectivement n). Supposons que $m \geq n$ et soit $\Phi : M \to N$ une application différentiable. On note $J\varphi$ le jacobien $\sqrt{\det({}^{\mathrm{t}}(D\Phi)(D\Phi))}$ de Φ. Pour $f : M \to \mathbb{R}^+$ une fonction mesurable positive, on a

$$\int_M f(x) J\Phi(x)\,\mathrm{d}\mathcal{H}_m(x) = \int_N \left(\int_{x \in \Phi^{-1}(\{y\})} f(x)\,\mathrm{d}\mathcal{H}_{m-n}(x) \right) \mathrm{d}\mathcal{H}_n(y)$$

Démonstration. Il existe une partition du complémentaire d'un ensemble (sous-analytique) de mesure 0 de M^n par des ouverts $V_1, \ldots, V_R$ telle que :

- L'ordre des simplexes de $\mathcal{K}(x)$ est constant sur chaque V_r ;
- Pour tout r et tout $x \in V_r$,
$$\mathcal{D}_s[\mathcal{K}(x)] = \sum_{i=1}^{N_r} \delta_{r_i},$$
où $r_i = (\varphi[J_{i_1}](x), \varphi[J_{i_2}](x))$ et N_r, J_{i_1}, J_{i_2} ne dépendent que de V_r ;
- J_{i_1} et J_{i_2} peuvent être choisis de sorte que
$$\Phi_{ir} : x \in V_r \to r_i = (\varphi[J_{i_1}](x), \varphi[J_{i_2}](x))$$
a rang maximal (2).

Le diagramme espéré peut donc s'écrire

$$\begin{aligned}\mathbb{E}\left[\mathcal{D}_s\left[\mathcal{K}\left(\mathbb{X}\right)\right]\right] &= \sum_{r=1}^{R} \mathbb{E}\left[\mathbb{1}\left\{\mathbb{X} \in V_r\right\} \mathcal{D}_s[\mathcal{K}(\mathbb{X})]\right] \\ &= \sum_{r=1}^{R} \mathbb{E}\left[\mathbb{1}\left\{\mathbb{X} \in V_r\right\} \sum_{i=1}^{N_r} \delta_{r_i}\right] \\ &= \sum_{r=1}^{R} \sum_{i=1}^{N_r} \underbrace{\mathbb{E}\left[\mathbb{1}\left\{\mathbb{X} \in V_r\right\} \delta_{r_i}\right]}_{=\mu_{i,r}}\end{aligned}$$

Par la formule de co-aire, en notant κ la densité de $\mathbb{X}$

$$\begin{aligned}\mu_{i,r}\left(B\right) &= P\left(\Phi_{i,r}\left(\mathbb{X}\right) \in B, \mathbb{X} \in V_r\right) \\ &= \int_{V_r} \mathbb{1}\left\{\Phi_{i,r}\left(x\right) \in B\right\} \kappa(x)\, \mathrm{d}\mathcal{H}_{nd}(x) \\ &= \int_{u \in B} \underbrace{\int_{x \in \Phi_{i,r}^{-1}(u)} \left(J\Phi_{i,r}\left(x\right)\right)^{-1} \kappa\left(x\right)\, \mathrm{d}\mathcal{H}_{nd-2}(x)}_{\text{Densité de } \mu_{i,r}} \mathrm{d}u.\end{aligned}$$

On conclut puisque la somme de mesure à densité a densité. ■

Si on revient à la Définition 3.16 de la surface de persistence de $D = \sum \delta_{r_i}$ avec noyau K, matrice de bande-passante H et poids w

$$\rho(D)(u) = \sum_i w(r_i) K_H(u - r_i) = D(wK_H(u - \cdot)),$$

on peut voir la surface de persistence comme un estimateur basé sur un noyau de $\mathbb{E}\left[\mathcal{D}_s\left[\mathcal{K}\left(\mathbb{X}\right)\right]\right]$.

4 Fonctions de Morse

4.1 Fonction Lisses

Dans la suite on considère des fonctions réelles sur une d-variété différentielle $\mathcal{M}$ munie d'un atlas.

Définition 4.1 Un *point critique* p de $f : \mathcal{M} \to \mathbb{R}$ est un point de $\mathcal{M}$ tel que $\nabla f(p) = 0$. Il est dit *dégénéré* si $|H(p)| = 0$.

Lemme 4.2 — Lemme de Morse. Si on se donne un point critique non-dégénéré p^* pour $f : \mathcal{M} \to \mathbb{R}$, il existe une carte φ de $\mathcal{M}$ telle que $\varphi(p^*) = 0$ et tel que sur l'ouvert associé à φ, f s'écrit

$$f(p) = f(p^*) - \sum_{i=1}^{q} x_i^2 + \sum_{i=q+1}^{d} x_i^2,$$

dans son développement de Taylor autour de p^*.

Définition 4.3 *L'indice* d'un point critique non-dégénéré p^* est le nombre q de coefficients négatifs du développement du Lemme de Morse. C'est le nombre de valeurs propres négatives de $H(p^*)$.

Définition 4.4 Une fonction $f : \mathcal{M} \to \mathbb{R}$ est une *fonction de Morse* si elle n'a pas de point critique dégénéré et des valeurs critiques distinctes.

Proposition 4.5
- Les fonctions de Morse forment un sous-ensemble dense des fonctions lisses.
- Les points critiques d'une fonction de Morse sont isolés (et f est localement polynomiale).

Dans la suite, on prend f une fonction de Morse et on note $\mathcal{M}^i = f^{-1}(]-\infty, i])$ l'ensemble de sous-niveau de f associé à i.

Proposition 4.6
- Si on se donne un intervalle $[i, j]$ qui ne contient pas de valeur critique de f, $\mathcal{M}^i$ et $\mathcal{M}^j$ sont difféomorphes.
- Autour d'une valeur critique $f(p)$ d'indice q, $\mathcal{M}^{f(p)-\varepsilon}$ auquel on recolle $\mathbb{D}^q$ a même homotopie que $\mathcal{M}^{f(p)+\varepsilon}$.

Théorème 4.7 — Relation de Morse-Euler On a toujours

$$\chi(\mathcal{M}) = \sum_{i=0}^{d} (-1)^i \mu_i(f),$$

ou $\mu_i(f)$ est le nombre de points critiques d'indice i de f.

4.2 Représentation des fonctions de Morse

Définition 4.8 Soit $p \in \mathbb{R}^n$ et σ un d-simplexe. Il existe des réels $\alpha_0, \dots, \alpha_d$ tels que

$$p = \sum_{i=0}^{d} \alpha_i \tau_0^i, \quad \sum_{i=0}^{d} \alpha_i = 1,$$

où τ_0^i est la i-ème 0-face de σ. Ces coefficients sont appelés *coordonnées barycentriques* de p relativement à σ.

Définition 4.9 Si $\hat{f}$ est une fonction réelle sur les 0-simplexes d'une triangulation $\mathcal{T}$, on définit $f : \mathcal{T} \to \mathbb{R}$ par interpolation linéaire de $\hat{f}$ par rapport aux coordonnées barycentriques, i.e. si $p \in \sigma \in \mathcal{T}$ est un point d'un d-simplexe,

$$f(p) = \sum_{i=0}^{d} \alpha_i \hat{f}(\tau_0^i).$$

f est un *champ scalaire linéaire par morceaux* (PL scalar field).

Définition 4.10 Le *lien inférieur* $\mathrm{Lk}^-(\sigma)$ (respectivement supérieur Lk^+) d'un d-simplexe σ relativement à un champ scalaire PL f est la partie du lien $\mathrm{Lk}(\sigma)$ telle que chacune de ses 0-faces a une image par f strictement inférieure (respectivement supérieure) à celle de σ.

Définition 4.11 Pour $f : \mathcal{M} \to \mathbb{R}$ un champ scalaire PL sur une variété PL $\mathcal{M}$, un sommet v de $\mathcal{M}$ est un *point régulier* si et seulement si $\mathrm{Lk}^-(v)$ et $\mathrm{Lk}^+(v)$ sont simplement connexes. Sinon, v est un *point critique* de f et $f(v)$ est une *isovaleur critique*.

Définition 4.12 Un champ scalaire PL f est une *fonction de Morse PL* si toutes ses isovaleurs critiques sont distinctes et si elle n'a pas de point critique dégénéré.

La relation de Morse-Euler nous dit que les points critiques se simplifient en paires. On va pour cela utiliser l'ordre associé à la persistence croissante pour simplifier les paires.

Définition 4.13 On construit des sous-complexes simpliciaux $\mathcal{M}_i$ de $\mathcal{M}$ une variété PL par l'union des i premiers simplexes pour l'ordre lexicographique. Ceci nous donne une suite de sous-complexes croissante pour l'inclusion : c'est une filtration appelée la *filtration lexicographique*.

Calculer la suite des groupes d'homologies $\mathcal{H}_p(\mathcal{M}_i)$ nous permet de calculer le diagramme de persistence associé à $\mathcal{M}$.

Définition 4.14 Si (τ, σ_j) est une paire de persistence (un élément d'un diagramme de persistence), on définit sa *persistence* comme $\tilde{f}(\sigma_j) - \tilde{f}(\tau)$ où pour tout simplexe σ, sa *valeur scalaire* $\tilde{f}(\sigma)$ est le maximum des valeurs de f sur les sommets (0-faces) de σ.

Ceci nous donne l'algorithme 3 pour calculer la persistence de chaque composante topologique de la variété PL $\mathcal{M}$.

Algorithme 3 Calcul des Paires de Persistence pour Morse

InputFiltration lexicographique de $\mathcal{M}$ par f
OutputDiagrammes de persistence $\text{diag}_k(f)$.
for $j \in [1, n]$ **do**
  // Traiter le $(d_i + 1)$-simplexe σ_j.
  $Pair(\sigma_j) \leftarrow \varnothing$
  $Chain(\sigma_j) \leftarrow \sigma_j$
  // Propagation homologue de $\partial\sigma_j$
  while $\partial\,(Chain(\sigma_j)) \neq 0$ **do**
    $\tau \leftarrow \max\,(\partial\,(Chain(\sigma_j)))$
    if $Pair(\tau) == \varnothing$ **then**
      // τ crée un (d_i)-cycle
      break
    else
      // Étendre la chaîne (avec bord homologue)
      $Chain(\sigma_j) \leftarrow Chain(\sigma_j) + Chain\,(Pair(\tau))$
  if $\partial\,(Chain(\sigma_j)) \neq 0$ **then**
    // Un cycle non-trivial homologue à $\partial\sigma_j$ existe (l.10)
    $\tau \leftarrow \max\,(\partial\,(Chain(\sigma_j)))$
    $Pair(\sigma_j) \leftarrow \tau$
    $Pair(\tau) \leftarrow \sigma_j$
    $\text{diag}_{d_i}(f) \leftarrow \text{diag}_{d_i}(f) \cup (\tau, \sigma_j)$

Cet algorithme suit la même idée que 1.

4.3 Comparaison Topologique

On a désormais une méthode pour calculer des diagrammes de persistence à partir uniquement d'une filtration et d'une fonction de Morse PL sur cette filtration.

4.3.1 Simplifcation et Empreinte Topologique

L'idée va donc être d'avoir une mesure permettant de simplifier la topologie de la variété PL en ne maintenant que les propriétés suffisamment persistentes :

$$\begin{array}{ccc} \mathcal{D}(f) & \longrightarrow & \mathcal{D}(g) \subseteq \mathcal{D}(f) \\ \uparrow & & \downarrow ? \\ f & & g \end{array}$$

La question est donc de savoir comment calculer la fonction g a partir du diagramme objectif $\mathcal{D}(g)$.

Toutefois, bien qu'en 2 dimensions le problème soit résolu, la simplification des ensembles de sous-niveau de paramètres (t, d) (calcul d'une fonction à distance au plus d de la fonction originelle qui minimise les nombres de Betti d'isovaleur t) est NP-difficile en 3 dimensions. Pour cela on va plutôt chercher à optimiser numériquement la fonction g, en se tournant vers l'analyse géométrique de l'espace des diagrammes de persistence.

Définition 4.15 On définit le diagramme $\mathcal{D}^*$ *barycentre* de N diagrammes d'entrée $\mathrm{diag}(f_i)$ comme

$$\mathcal{D}^* = \mathrm{argmin}_{\mathcal{D}} \sum_{i=1}^{N} W_2^2(\mathcal{D}, \mathrm{diag}(f_i)).$$

On rappelle que le Théorème 2.10 montre que si f_i et f_j sont proches, leurs diagrammes le seront.

En voyant les diagrammes de persistence comme des empreintes topologiques des variétés. En particulier, on peut comparer les données, notamment via la distance de Wasserstein qui se calcule en résolvant un simple problème d'assignation, et ceci permet de définir des géodésiques et des moyennes sur entre les variétés, ce qui permet de construire de meilleurs groupement avec l'algorithme des k-moyennes par exemple.

4.3.2 Graphes de Reeb

Les diagrammes de persistence sont malheureusement assez limités, notamment puisque f et λf auront toujours le même diagramme associé, ou par exemple si on échange les centres de gaussiennes dans une mixture.

Définition 4.16 Si $f : \mathcal{M} \to \mathbb{R}$ est un champ scalaire de Morse PL défini sur une variété compacte PL $\mathcal{M}$, on note $f^{-1}(f(p))_p$ le *contour* de f contenant $p \in \mathcal{M}$. Le *graphe de Reeb* $\mathcal{R}(f)$ est le complexe simplicial unidimensionnel défini comme le quotient de $\mathcal{M} \times \mathbb{R}$ par $(p_1, f(p_1)) \sim (p_2, f(p_2)) \Leftrightarrow p_2 \in (f^{-1}(f(p_1)))_p \wedge f(p_1) = f(p_2)$.

Le graphe de Reeb est calculable en temps $\mathcal{O}(m \log m)$ où m est le nombre de 2-simplexes de $\mathcal{M}$.

Proposition 4.17 Si on note $\varphi : \mathcal{M} \to \mathcal{R}(f)$ la projection de la variété sur le graphe de Reeb associé à f et $\psi : \mathcal{R}(f) \to \mathbb{R}$ la projection sur l'isovaleur associée au point, on a $f = \psi \circ \varphi$.

Définition 4.18 La valence d'un 0-simplexe $v \in \mathcal{R}(f)$ est le nombre de 1-simplexes dans son étoile $\mathrm{St}(v)$.

Proposition 4.19 Soit $f = \psi \circ \varphi$ un champ scalaire de Morse PL sur une d-variété PL et soit $\mathcal{R}(f)$ son graphe de Reeb.

- On caractérise les images par φ des points réguliers de f comme l'intérieur de 1-simplexes de $\mathcal{R}(f)$.
- On caractérise les images des points critiques d'indice 0 ou d de f (donc ses extrema) par les 0-simplexes de $\mathcal{R}(f)$ de valence 1.
- Si $d = 2$, on caractérise les images des points critiques d'indice 1 de f (donc ses selles) comme les 0-simplexes de $\mathcal{R}(f)$ de valence $2, 3, 4$.
- Si $d \geq 3$, les points critiques d'indice 1 ou $d-1$ de f ont pour image par φ les 0-simplexes de valence 2 ou 3. La réciproque n'est pas nécessairement vraie.
- Si $d > 3$, les points critiques d'indice différent de $0, 1, (d-1)$ ou d ont pour image par φ les 0-simplexes de valence 2.

En terme d'homologie, il reste à comparer l'homologie et les graphes de Reeb.

Théorème 4.20 On a $\beta_0(\mathcal{R}(f)) = \beta_0(\mathcal{M})$ et $\beta_1(\mathcal{R}(f)) \leq \beta_1(\mathcal{M})$.

On peut noter que dans le cas où $\mathcal{M}$ est une surface (2-variété), le graphe de Reeb est planaire.

Théorème 4.21 On considère $\mathcal{R}(f)$ le graphe de Reeb de f un champ scalaire PL sur une 2-variété $\mathcal{M}$. Soit $b(\mathcal{M})$ le nombre de composantes connexes de $\mathcal{M}$ et $g(\mathcal{M})$ son genre. Le nombre de boucles de $\mathcal{R}(f)$ est décrit par :

- Si $\mathcal{M}$ est orientable :
  - Si $b(\mathcal{M}) = 0$ alors $\beta_1(\mathcal{R}(f)) = g(\mathcal{M})$
  - Sinon $g(\mathcal{M}) \leq \beta_1(\mathcal{R}(f)) \leq 2g(\mathcal{M}) + b(\mathcal{M}) - 1$
- Sinon :
  - Si $b(\mathcal{M}) = 0$ alors $0 \leq \beta_1(\mathcal{R}(f)) \leq \frac{1}{2}g(\mathcal{M})$
  - Sinon $0 \leq \beta_1(\mathcal{R}(f)) \leq g(\mathcal{M}) + b(\mathcal{M}) - 1$

Définition 4.22 Étant donnée $f : \mathcal{M} \to \mathbb{R}$, on note $x \sim_M y$ si $f(x) = f(y) = \alpha$ et x, y sont dans la même composante connexe de $f^{-1}(]-\infty, \alpha])$. L'*arbre de fusion* est l'espace $\mathcal{M}$ quotienté par $\sim_M$.

L'arbre de fusion est une relaxation du graphe de Reeb et permet de supprimer les boucles du graphe de Reeb en supprimant la possibilité de défusionner. Il permet de plus de récupérer un historique des fusions de composantes.

Le graphe de Reeb permet de faire de la segmentation par arcs en considérant les séparations et les fusions comme des changements de segments. Toutefois, ils perdent beaucoup d'information, on introduit donc le cartographe :

Définition 4.23 Si $f : \mathcal{M} \to \mathbb{R}$ tire en arrière tous les intervalles en un nombre fini de composantes connexes par arc, on définit le cartographe $M(I, f)$ (ou mapper) de l'intervalle I en f comme le nerf de la famille des composantes connexes de $f^{-1}(I)$.

Dans le cas d'un nuage de points, on remplace les composantes connexes par les clusters sous f^{-1}, et on obtient un algorithme qui permet de calculer localement une alternative plus générale au graphe de Reeb.

5 Théorie de Morse Discrète

5.1 Complexe de Morse

6 Inférence Topologique