Journées mathématiques X-UPS
2024, p. 1–20 doi: 10.5802/xups.2024-01

INTRODUCTION À LA THÉORIE DE LA PERSISTANCE À TRAVERS UN EXEMPLE D'APPLICATION

par

Steve Oudot

Résumé. Dans ce texte nous présentons de manière informelle les idées qui sous-tendent la théorie de la persistance, en nous appuyant sur un exemple particulier d'application de cette théorie au regroupement de données.

Table des matières

Le but de ce texte est d'introduire quelques unes des principales idées sur lesquelles repose la théorie de la persistance, qui est le fondement mathématique de l'analyse topologique de données et le sujet central de cet ouvrage. Pour ce faire, nous allons nous placer dans un cadre restreint, celui de la persistance des pics d'une fonction réelle, et nous allons prendre pour prétexte l'application de la théorie à la classification non supervisée (aussi appelée *regroupement*) en apprentissage machine. L'exposé alternera donc entre des sections de présentation du contexte applicatif et des sections plus formelles introduisant les notions mathématiques. Parmi ces sections, la seule qui soit vraiment utile pour les textes suivants est la section 4, qui

introduit la persistance dans notre cadre. Le lecteur qui ne s'intéresserait pas au contexte applicatif peut sans risque limiter sa lecture à cette seule section avant de passer au texte suivant.

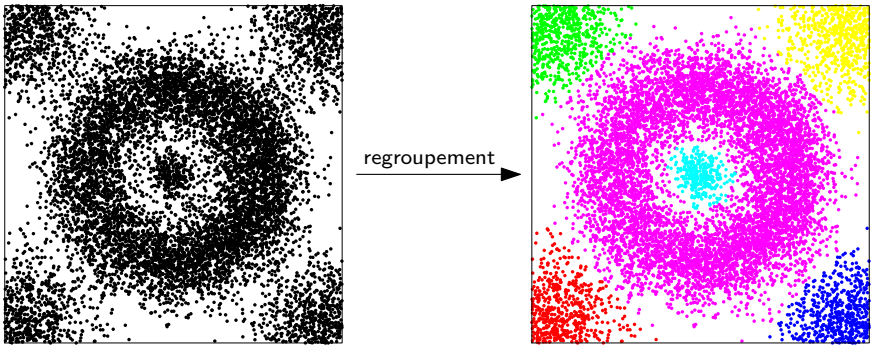


FIGURE 1. Un échantillon de points dans $\mathbb{R}^2$ (à gauche) et un regroupement de ces points (à droite).

1. Le problème du regroupement

Soit $P = \{p_1, \ldots, p_n\}$ un ensemble fini de points de l'espace euclidien $\mathbb{R}^d$ – dans le jargon on parle d'*échantillon de points*, de *jeu de données* ou encore de *nuage de points*, selon le domaine scientifique considéré (statistiques, apprentissage machine, ou géométrie). On suppose que les points proviennent de m classes distinctes, que l'on ne connaît pas – m lui-même n'est généralement pas connu non plus. L'objectif est donc (au besoin) de déterminer le nombre m de classes, puis de partitionner le nuage P en m sous-ensembles appelés *clusters*, $P = \bigsqcup_{\ell=1}^{m} C_\ell$, de manière à respecter au mieux les m classes sous-jacentes. Voir la figure 1 pour une illustration. Cette partition du nuage P s'appelle un *regroupement* des points (*clustering* en anglais).

Tel que formulé, le problème est clairement mal posé puisque, en l'absence d'informations complémentaires sur la manière dont les points sont obtenus à partir des classes sous-jacentes, il n'y a aucun moyen de déterminer si un regroupement du nuage est meilleur qu'un autre. Il faut donc formuler des *hypothèses sur la génération des données*.

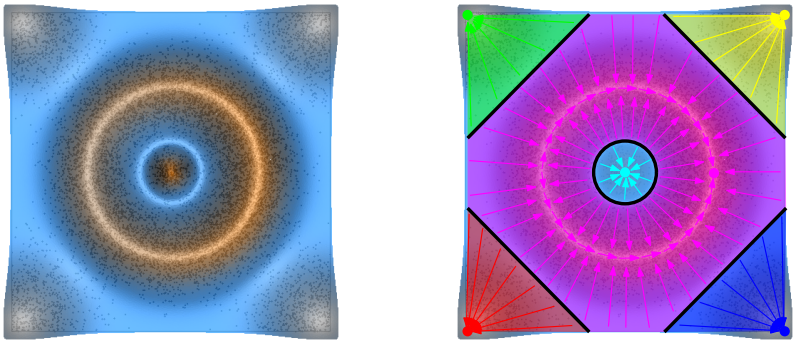


FIGURE 2. À gauche : le nuage de points P de la figure 1 superposé à la densité de probabilité f selon laquelle les points de P ont été échantillonnés iid. Les valeurs de f sont représentées par une palette de couleurs, du bleu (densité faible) à orange (densité forte). À droite : les six pics de f et leurs variétés stables (de la même couleur que le pic correspondant), ainsi qu'une représentation synthétique du flot de gradient de f.

Chaque méthode de clustering vient avec son propre jeu d'hypothèses, duquel découle un critère de qualité sur les partitions de P. Ici, on va s'intéresser aux approches dites *basées sur la densité*, qui formulent l'hypothèse suivante qu'on adoptera tout au long de ce texte – voir la figure 2 (gauche) pour une illustration :

Hypothèse 1.1. Les points $p_1, \ldots, p_n$ sont échantillonnés de manière iid selon une mesure de probabilité μ, de densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$. La mesure μ comme sa densité f sont supposées *inconnues*.

Les classes sont alors définies comme des sous-ensembles deux à deux disjoints de $\mathbb{R}^d$ associés aux maxima locaux – aussi appelés *modes* ou *pics* – de la densité f. Plus précisément, elles correspondent aux *variétés stables* des modes, c'est-à-dire qu'il y a exactement une classe par mode x, et cette classe est définie comme étant le lieu des points de $\mathbb{R}^d$ qui convergent asymptotiquement vers x lorsqu'ils sont poussés continûment le long du flot de gradient de la fonction f –

voir la figure 2 (droite) pour une illustration. Nous allons maintenant formaliser cette notion en utilisant un peu de théorie de Morse, dont une référence classique est le livre de Milnor [Mil63].

2. Définition mathématique des classes à partir de la densité f

Comme on vient de le dire, pour définir les classes associées aux modes de notre densité f nous allons pousser les points de l'espace $\mathbb{R}^d$ continûment le long du flot de gradient de f. De toute évidence il nous faut faire des hypothèses de régularité sur f afin d'avoir un gradient et un flot bien définis, ainsi qu'un contrôle sur la convergence des trajectoires des points. Voici un jeu simplifié d'hypothèses, qu'on supposera vérifiées dans toute la suite du texte :

Hypothèse 2.1. On suppose que la densité f :

(i) est lisse, de classe C^∞ ;
(ii) s'annule à l'infini, c'est-à-dire que $\lim_{\|x\|\to\infty} f(x) = 0$;
(iii) est *de type Morse*, c'est-à-dire que ses *points critiques*, i.e., les points x où le gradient $\nabla f(x)$ s'annule, sont en nombre fini et *non dégénérés*, c'est-à-dire que la matrice hessienne de f en chacun de ces points est inversible.

L'hypothèse (iii) joue un rôle clé dans la suite. Elle implique en particulier que les points critiques de f sont isolés dans $\mathbb{R}^d$. Bien qu'elle puisse paraître plus restrictive que les autres hypothèses a priori, puisqu'elle impose des conditions sur les quantités différentielles d'ordre 1 et 2 de f, il s'avère que les fonctions de type Morse forment un ouvert dense pour la topologie C^2 dans l'espace des fonctions C^∞, donc on ne perd presque rien en généralité en ajoutant l'hypothèse (iii).

Comme la fonction f est de classe C^∞, son champ de gradient est localement lipschitzien et peut donc être intégré en un flot continu $\Phi\colon \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^d$. Plus précisément, par le théorème de Cauchy-Lipschitz global, la ligne de flot $\varphi_x\colon t \mapsto \Phi(t,x)$ issue d'un point $x \in \mathbb{R}^d$

est l'unique solution de l'équation différentielle ordinaire suivante :

$$\begin{cases} \dot{\varphi}_x(t) = \nabla f(\varphi_x(t)) \\ \varphi_x(0) = x \end{cases}$$

Cette solution dépend continûment à la fois du paramètre t et de la condition initiale x, d'où la continuité du flot Φ.

On regarde maintenant, pour tout point $x \in \mathbb{R}^d$, si et où converge la ligne de flot φ_x lorsque $t \to +\infty$. Le cas particulier où $f(x) = 0$ est trivial : en tant que densité de probabilité, f est positive ou nulle, donc x est un minimum local de f et, à ce titre, lui-même un point critique de f, stationnaire pour le flot Φ par définition. Pour le cas général $f(x) > 0$, l'hypothèse que f s'annule à l'infini (sur $\mathbb{R}^d$ qui est localement compact) implique que l'ensemble de sur-niveau

$$\{y \in \mathbb{R}^d \mid f(y) \geqslant f(x) > 0\},$$

dans lequel se trouve l'image de la ligne de flot φ_x, est compact ; il s'ensuit alors que la ligne de flot φ_x converge bien lorsque $t \to +\infty$, et par définition la limite est un point critique de f. Ainsi, à la limite, le flot amène tous les points de $\mathbb{R}^d$ à des points critiques de f.

Définition 2.2. La *variété stable* d'un point critique x de f est l'ensemble des points $y \in \mathbb{R}^d$ tels que $x = \lim_{t\to+\infty} \varphi_y(t)$.

Il découle de ce qui vient d'être dit que les variétés stables des points critiques partitionnent l'espace $\mathbb{R}^d$. Toutefois, tous les points critiques ne sont pas des pics : il y a également les minima locaux, ainsi que les points critiques de type *selle*. Dans la suite on ne retiendra que les variétés stables des pics.

Définition 2.3. Les classes correspondant à la densité f sont par définition les variétés stables des pics de f.

L'idée derrière le fait de ne regarder que les variétés stables des pics est que, en général, identifier les points critiques requiert d'évaluer des quantités différentielles, typiquement le gradient de f (voire la matrice hessienne si on veut déterminer le type des points critiques), dont le calcul est instable numériquement. Les pics, quant à eux, ne nécessitent pas de quantités différentielles pour être caractérisés

(définition 4.1), leur calcul en pratique s'avère donc beaucoup plus stable numériquement.

En contrepartie, on peut s'interroger sur la pertinence de laisser de côté les autres types de points critiques, impliquant de fait que l'union des classes ne couvre pas $\mathbb{R}^d$ tout entier. En réalité, la théorie de Morse nous garantit que les classes couvrent presque tout l'espace :

Proposition 2.4. *Sous l'hypothèse* 2.1*, le complémentaire de l'union des variétés stables des pics de f est une union finie de sous-variétés différentielles de $\mathbb{R}^d$ de codimensions strictement positives, et donc de mesure de Lebesgue nulle dans $\mathbb{R}^d$.*

En pratique, la probabilité qu'un point p_i de notre échantillon P ne soit pas parmi les classes est donc nulle, puisque les p_i sont échantillonnés selon la mesure μ qui a une densité par rapport à la mesure de Lebesgue. Ainsi, nos hypothèses initiales et notre définition des classes sont génériquement compatibles avec notre problème.

3. Calcul des clusters à partir du nuage de points P

Comme on l'a supposé dans l'hypothèse de départ 1.1, la mesure μ et sa densité f ne sont pas connues en pratique. Il nous va donc falloir trouver un moyen de simuler la montée de gradient à partir des données $p_1, \ldots, p_n$ pour pouvoir former des clusters qui approximent les classes. Il existe tout un éventail de méthodes pour ce faire. Les unes, comme par exemple *mean-shift* [CM02], adoptent une approche purement numérique en tentant d'approximer localement le gradient de f puis de simuler la montée de gradient continue par une montée de gradient approximé discrète dans $\mathbb{R}^d$. Les autres, plus anciennes comme celle que nous allons voir ici [KNF76], adoptent une approche combinatoire en remplaçant $\mathbb{R}^d$ par un objet discret appelé *graphe de voisinage*, construit à partir des données, dans lequel le gradient est approximé en chaque sommet par une arête incidente.

Pré-traitement. En pré-traitement on construit le graphe de voisinage et on approxime la densité aux sommets du graphe.

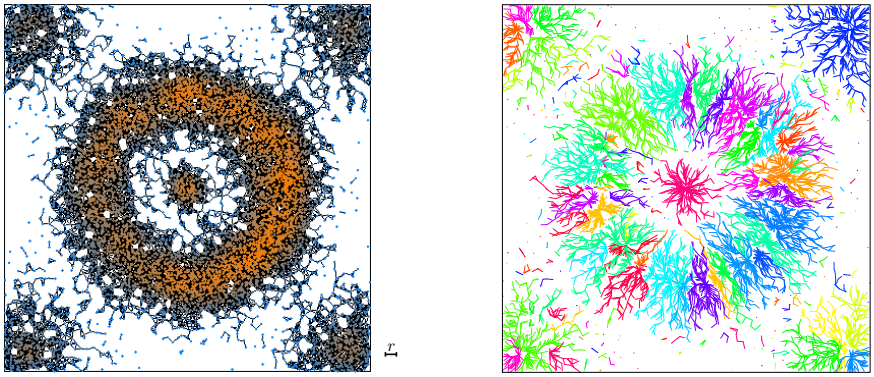


FIGURE 3. À gauche : le graphe G de r-voisinage sur le nuage de points P de la figure 1, pour la valeur de r montrée au centre. Les arêtes du graphe sont en noir, tandis que les valeurs de l'estimateur $\widehat{f}$ aux sommets sont représentées comme à la figure 2. À droite : le résultat de l'algorithme sur cette entrée, avec une couleur distincte par cluster. Les arêtes représentées sur le dessin sont celles des arbres de la forêt couvrante de G calculée par l'algorithme.

Définition 3.1. Un *prédicat de voisinage* est une fonction symétrique $P \times P \to \{0,1\}$ qui vaut 0 sur la diagonale $\{(p_i, p_i) \mid p_i \in P\}$.

Définition 3.2. Étant donné un prédicat de voisinage $\sigma\colon P \times P \to \{0,1\}$, le *graphe de voisinage* correspondant est le graphe combinatoire $G = (P, E)$ non orienté dont les sommets sont les points de P et les arêtes forment l'ensemble $E = \{(p_i, p_j) \in P \times P \mid \sigma(p_i, p_j) = 1\}$.

Exemple 3.3. Voici deux constructions classiques de graphes de voisinage, dont la première est illustrée dans la figure 3 (gauche) :

- le graphe de *r-voisinage*, pour un réel $r \geqslant 0$, correspond au prédicat $(p_i, p_j) \mapsto \mathbb{1}_{p_i \neq p_j}\, \mathbb{1}_{\|p_i - p_j\|_2 \leqslant r}$ qui teste si deux points donnés sont à distance euclidienne au plus r l'un de l'autre ;
- le graphe de *k-voisinage*, pour un entier $k \geqslant 1$, correspond au prédicat $(p_i, p_j) \mapsto \mathbb{1}_{p_i \neq p_j}\, \mathbb{1}_{p_j \in \mathrm{ppv}_k(p_i)}\, \mathbb{1}_{p_i \in \mathrm{ppv}_k(p_j)}$, où $\mathrm{ppv}_k(p_\ell)$ désigne l'ensemble des k plus proches voisins de p_ℓ parmi les points du nuage P pour la distance euclidienne.

Une fois un tel graphe de voisinage $G = (P, E)$ construit à partir de P, on calcule une estimation $\widehat{f}(p_i)$ de la densité f en chaque point $p_i \in P$. L'estimation de la densité est en soi un sujet à part entière, qui sort du cadre de cet exposé. Notons simplement que le domaine des statistiques nous fournit tout un éventail d'estimateurs ayant chacun des propriétés spécifiques. Ici nous ferons simplement l'hypothèse (relativement forte mais classique) que l'estimateur $\widehat{f}$ approxime la densité f en norme sup sur P, c'est-à-dire que l'on peut borner l'erreur de l'estimateur comme suit :

$$\|\widehat{f} - f\|_\infty \stackrel{\text{déf}}{=} \max_{1 \leqslant i \leqslant n} |\widehat{f}(p_i) - f(p_i)| \leqslant \varepsilon(n) \tag{3.1}$$

où la quantité $\varepsilon(n)$ dépend uniquement de n, pas des points du nuage P, et tend vers 0 lorsque $n \to +\infty$. La borne $\varepsilon(n)$ en elle-même n'a pas d'importance ici, et pour simplifier nous la supposerons vérifiée de manière déterministe et non pas seulement avec forte probabilité comme c'est le cas en pratique. Ainsi nous avons équipé notre graphe de voisinage $G = (P, E)$ d'un champ scalaire $\widehat{f} \colon P \to \mathbb{R}^+$ approximant la densité f aux sommets – voir encore la figure 3 (gauche). C'est l'entrée que nous fournissons à l'algorithme.

L'algorithme. En chaque sommet p_i du graphe nous approximons le gradient de f par l'arête reliant p_i à son voisin p_j dont l'estimée $\widehat{f}(p_j)$ est la plus élevée, *à condition que celle-ci soit plus élevée que* $\widehat{f}(p_i)$. Dans le cas contraire, p_i est un maximum local de $\widehat{f}$ dans le graphe G et on le déclare donc comme étant un pic de la densité.

L'ensemble des arêtes ainsi sélectionnées forme une *forêt couvrante* de G, c'est-à-dire un ensemble de sous-graphes qui sont des arbres (i.e., des sous-graphes connexes sans cycles) et qui couvrent des sous-ensembles deux à deux disjoints de sommets dont l'union est le nuage P tout entier[(1)]. Ces arbres sont les clusters produits par l'algorithme. Chacun contient une unique racine, son sommet dont la valeur de $\widehat{f}$ est la plus élevée, qui par construction est un maximum local de $\widehat{f}$ dans G et sert donc d'approximation pour un éventuel

[(1)]Techniquement, certains sommets de G peuvent être isolés, auquel cas on les ajoute à la forêt en tant qu'arbres singletons.

pic de f. L'arbre en lui-même sert d'approximation pour la variété stable associée à ce pic dans $\mathbb{R}^d$.

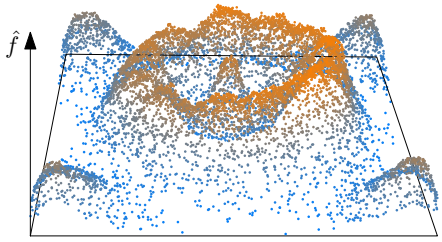


FIGURE 4. Graphe de l'estimateur de densité $\widehat{f}$ restreint au nuage de points P de la figure 1.

Résultat. La figure 3 (droite) montre un exemple de résultat de l'algorithme, qui, il faut l'avouer, est de piètre qualité. Ce qui frappe immédiatement, c'est la multiplication des clusters (plusieurs dizaines) par rapport au nombre de classes sous-jacentes (six seulement). Ceci s'explique essentiellement par le fait que l'estimateur $\widehat{f}$ est beaucoup plus bruité que la densité f, n'étant qu'une approximation en norme sup d'après (3.1) – voir la figure 4 pour une illustration du bruit dans l'estimateur. Ainsi, en plus de quelques pics « légitimes » associés aux pics de f dans $\mathbb{R}^d$, s'ajoute dans le graphe G tout un tas de pics fallacieux dus au bruit dans $\widehat{f}$. Afin de distinguer parmi les pics ceux qui sont légitimes de ceux qui ne le sont pas, nous allons utiliser une notion mathématique qui quantifie l'importance de chaque pic : la *persistance*. Cette notion va également nous fournir une *hiérarchie sur les pics* de $\widehat{f}$ dans G, hiérarchie qui va nous permettre de réparer le clustering produit par l'algorithme en fusionnant les clusters associés aux pics fallacieux avec les clusters de leurs parents dans la hiérarchie.

4. La persistance des pics d'une fonction réelle

Alors que le terme *persistance* est utilisé communément en analyse topologique de données, dans le cas particulier des pics il renvoie à un concept plus ancien, la *proéminence*, introduit par les alpinistes et

les géographes, que nous allons présenter en premier afin de nourrir l'intuition du lecteur.

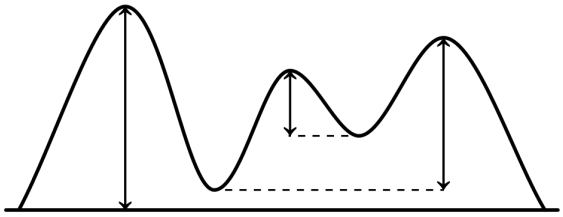


FIGURE 5. Les flèches verticales montrent la proéminence de trois pics sur une île. Une ligne pointillée horizontale relie chaque pic (excepté le plus haut) à son col le plus élevé. Source : *Wikipedia* (`https://en.wikipedia.org/wiki/File:Relative-height.png`, image sous licence CC BY-SA 3.0 Deed).

4.1. Définition des alpinistes et géographes. Avec[(2)] le développement de l'exploration alpine dans la deuxième moitié du XIX^{e} siècle, des listes de sommets, conquis ou à conquérir, ont commencé à émerger. Rapidement s'est posée la question de déterminer ce qu'est un sommet. En effet, sans critère restrictif, n'importe quelle antécime, épaule, ou même pierre pourrait être vue comme un sommet en elle-même. Le critère principal retenu pour distinguer les sommets des autres types de protubérances a été celui de la proéminence, qui doit être supérieure à un certain seuil pour que la protubérance puisse être considérée comme un sommet indépendant. Ce seuil varie d'un classement à l'autre : de 30 mètres (la longueur de corde de l'alpinisme classique) pour la liste officielle des 82 sommets des Alpes de plus de 4000 mètres, à 1500 mètres pour les sommets dits *ultra-proéminents*.

La proéminence pour les alpinistes et les géographes se définit comme « la différence d'altitude entre un sommet donné et l'ensellement ou le col le plus élevé permettant d'atteindre une cime encore plus haute ». Autrement dit, c'est « le dénivelé minimum de la descente à parcourir pour pouvoir remonter sur un sommet plus élevé ».

[(2)]Le contenu de cette sous-section est largement repris de l'article correspondant sur *Wikipedia* (`fr.wikipedia.org/wiki/Proéminence`).

TABLE 1. Liste des 10 sommets les plus hauts du monde, classés par ordre décroissant d'altitude. Le seuil minimal de proéminence requis ici est de 500 mètres.

sommet	continent	altitude (m)	proéminence (m)
Everest	Asie	8849	8849
K2	Asie	8611	4020
Kangchenjunga	Asie	8586	3922
Lhotse	Asie	8516	610
Makalu	Asie	8485	2378
Cho Oyu	Asie	8188	2340
Dhaulagiri I	Asie	8167	3357
Manaslu	Asie	8163	3092
Nanga Parbat	Asie	8126	4608
Annapurna I	Asie	8091	2984

TABLE 2. Liste des 10 sommets les plus proéminents du monde, classés par ordre décroissant de proéminence.

sommet	continent	altitude (m)	proéminence (m)
Everest	Asie	8849	8849
Aconcagua	Amérique	6960	6960
Denali	Amérique	6191	6155
Kilimanjaro	Afrique	5895	5885
Pico Cristóbal Colón	Amérique	5700	5509
Mont Logan	Amérique	5959	5250
Pico de Orizaba	Amérique	5636	4922
Massif Vinson	Antarctique	4892	4892
Puncak Jaya	Océanie	4884	4884
Mont Elbrouz	Europe	5642	4741

Voir la figure 5 pour une illustration. Notons que l'ensellement peut se situer au niveau de la mer mais pas au-dessous. Ainsi, le plus haut pic d'une île a une proéminence égale à sa hauteur. Les tables 1 et 2 donnent les listes des dix sommets les plus hauts du monde d'une part, des dix sommets les plus proéminents du monde d'autre part, et comme on peut le constater elles sont bien différentes.

4.2. Définition mathématique. On va maintenant définir formellement la proéminence, ou persistance. Pour cela on fixe un espace topologique X et une fonction $f\colon X \to \mathbb{R}$, sans plus d'hypothèses pour le moment – des hypothèses sur f et X apparaîtront au cours de l'exposé. Notons dès à présent que l'approche adoptée ici pour définir la persistance transcrit directement les idées de la section 4.1 en requérant peu d'outils mathématiques. En contrepartie, elle donne lieu à des énoncés parfois inutilement techniques et se généralise mal – voir à ce propos l'hypothèse 4.9 et les exemples et commentaires qui l'entourent. La bonne approche, fondée sur l'homologie, sera adoptée dans le texte [Oud24] – voir en particulier la remarque 5.11.

Définition 4.1. Un point $x \in X$ est un *maximum local* (ou *pic*) de f s'il existe un voisinage U de x dans X tel que $f(x) = \max_U f$.

Pour quantifier la proéminence des pics, on regarde l'évolution des composantes connexes par arc dans les *sur-niveaux* de la fonction f alors que le niveau diminue progressivement de $+\infty$ jusqu'à $-\infty$.

Définition 4.2. Étant donné un niveau $t \in \mathbb{R}$, l'*(ensemble de) sur-niveau* de f associé est $f^{-1}([t, +\infty))$.

Définition 4.3. Étant donné un pic $x \in X$ de f, pour tout niveau $t \leqslant f(x)$ on définit $C_t(x)$ comme étant la composante connexe par arc du sur-niveau $f^{-1}([t, +\infty))$ à laquelle appartient x.

Lorsque t diminue, le sur-niveau de f associé ne fait que croître, ainsi que ses composantes connexes par arc. En conséquence :

Corollaire 4.4. *Étant donné un pic $x \in X$ de f, pour tous niveaux $t \leqslant t' \leqslant f(x)$ on a $C_t(x) \supseteq C_{t'}(x)$.*

Ainsi, informellement, on peut traquer la croissance de la composante connexe par arc contenant notre pic x tandis que l'on abaisse le niveau t, et détecter la première valeur de t à laquelle cette composante fusionne avec celle d'un pic plus élevé : à cette valeur $t = h(x)$ particulière, x émerge comme un pic secondaire d'une montagne plus élevée, et sa persistance est donnée par la différence $f(x) - h(x) \geqslant 0$. Dans le cas particulier où la composante de x fusionne avec celle d'un

pic de même hauteur, on départage les deux pics en désignant l'un comme étant secondaire de l'autre de manière arbitraire.

Formellement, on suppose donné un ordre total $\preccurlyeq$ sur X (ce qui en toute généralité requiert une version faible de l'axiome du choix) et on considère l'ordre lexicographique suivant sur X :

$$y \geqslant x \iff \begin{cases} f(y) > f(x) & \text{ou} \\ f(y) = f(x) & \text{et} \quad y \preccurlyeq x \end{cases} \tag{4.1}$$

On note $>$ l'ordre total strict associé :

$$y > x \iff y \geqslant x \quad \text{et} \quad y \neq x$$

Définition 4.5. Pour tout pic $x \in X$ de f on définit :

- l'instant de *naissance* de x comme étant la valeur $f(x) \in \mathbb{R}$;
- l'instant de *décès* de x par $h(x) = \sup I(x) \in \mathbb{R} \cup \{-\infty\}$, où $I(x) = \{t \leqslant f(x) \mid \exists y > x \text{ pic de } f \text{ tel que } C_t(y) = C_t(x)\}$;
- l'*intervalle de persistance* de x comme étant l'intervalle semi-ouvert $]h(x), f(x)] \subseteq \mathbb{R}$, que l'on maintient ouvert à gauche par convention car cette extrémité peut être à l'infini ;
- la *persistance* (ou *proéminence* en géographie) de x comme étant la différence $f(x) - h(x) \in \mathbb{R}^+ \cup \{+\infty\}$.

Remarque 4.6. Pour les pics x tels que $h(x) = -\infty$, la proéminence telle que définie ici est infinie, tandis que pour les géographes la proéminence de ces pics est égale à leur hauteur (voir la section 4.1 et en particulier la figure 5).

Définition 4.7. Le *code-barres* de f est le multi-ensemble des intervalles de persistance des pics de f. La *multiplicité* d'un intervalle est le nombre (potentiellement infini) de ses occurrences dans le multi-ensemble.

Exemple 4.8. Considérons la fonction $x \mapsto -\sin(x)\cos(3x)$ sur l'intervalle $X = [-5, 2] \subset \mathbb{R}$ muni de l'ordre usuel sur les réels, dont le graphe est représenté à la figure 6. Cette fonction possède cinq pics, aux abscisses $x = -5$, $x = \arctan\sqrt{2 + \sqrt{11/3}} + k\,\pi$ et

FIGURE 6. Graphe (en brun) et code-barres (en violet) de la fonction $x \mapsto -\sin(x)\cos(3x)$ sur l'intervalle $X = [-5, 2]$ muni de l'ordre usuel sur les réels. Par souci de clarté, les intervalles de persistance sont ancrés à l'aplomb des pics correspondants de la fonction. Les liens de parenté formant la hiérarchie des pics sont matérialisés par des flèches courbes vertes reliant chaque pic à son parent lorsqu'il existe.

$x = -\arctan\sqrt{2-\sqrt{11/3}} + k\,\pi$ pour $k \in \{-1,\, 0\}$. Les intervalles de son code-barres sont, de gauche à droite :

$$]-a,\, c] \qquad]-b,\, b] \qquad]-\infty,\, a] \qquad]-b,\, b] \qquad]-a,\, a]$$

où

$$a = \left(3\sqrt{11/3}+5\right)\left(2+\sqrt{11/3}\right)^{1/2}\left(3+\sqrt{11/3}\right)^{-2} \;\approx\; 0.88$$
$$b = \left(3\sqrt{11/3}-5\right)\left(2-\sqrt{11/3}\right)^{1/2}\left(3-\sqrt{11/3}\right)^{-2} \;\approx\; 0.19$$
$$c = \sin(5)\cos(15) \;\approx\; 0.73$$

Les notions d'intervalle de persistance et de code-barres sont définies pour toute fonction $f\colon X \to \mathbb{R}$, toutefois elles n'ont de sens que lorsqu'on fait l'hypothèse que les composantes connexes par arc

contenant les pics couvrent l'intégralité des sur-niveaux de la fonction, c'est-à-dire :

Hypothèse 4.9. $\forall t \in \mathbb{R}, \quad f^{-1}([t, +\infty[) = \bigcup_{\substack{x \text{ pic de } f \\ f(x) \geqslant t}} C_t(x)$

Dans le cas contraire en effet, des composantes connexes par arc dans les sur-niveaux peuvent être ignorées et donc des barres ne pas apparaître ou bien apparaître avec des extrémités erronées dans le code-barres, comme dans les exemples ci-dessous :

Exemple 4.10. La fonction identité sur $\mathbb{R}$ ou sur l'intervalle $]0, 1[$ n'a aucun pic et donc son code-barres est vide. De même pour la fonction

$$x \longmapsto \begin{cases} 1 - |x| & \text{si } x \neq 0 \\ 0 & \text{si } x = 0 \end{cases} \quad \text{sur l'intervalle } [-1, 1].$$

Exemple 4.11. La fonction $f\colon x \mapsto -(x^3 + 2)\exp(-x)$ sur $\mathbb{R}^+$ a un unique pic en $x = 1$, dont la proéminence est $+\infty$ alors qu'on s'attendrait à ce qu'elle soit finie car $f(1) = -3/e$ et $\lim_{t\to+\infty} f = 0$.

À partir de maintenant nous supposerons donc l'hypothèse 4.9 vérifiée. C'est le cas par exemple lorsque X est compact et f est continue, ou encore lorsque $X = \mathbb{R}^d$ et f est continue, positive ou nulle et s'annule à l'infini comme sous l'hypothèse 2.1.

Hiérarchie des pics. En plus du code-barres, nous pouvons définir une notion de parent et de là une hiérarchie sur les pics. Pour cela nous faisons l'hypothèse additionnelle suivante, également vérifiée sous l'hypothèse 2.1 :

Hypothèse 4.12. Le nombre de pics de la fonction f est fini.

Exercice 4.13. Soit $x \in X$ un pic de f dont la proéminence est finie ($h(x) > -\infty$). Montrer que l'ensemble $I(x)$ de la définition 4.5 est alors non vide, de même que l'ensemble

$$J(x) \;=\; \{y \text{ pic de } f \mid y > x \text{ et } C_t(y) = C_t(x)\ \forall t \in I(x)\}$$

et que ce dernier admet un maximum pour l'ordre $\geqslant$ de l'équation (4.1).

Le maximum de $J(x)$ est appelé le *parent* de x. Il est strictement supérieur à x pour l'ordre $\geqslant$ donc la relation de parenté induit une hiérarchie sur les pics, au sommet de laquelle se trouvent les pics de proéminence infinie. La figure 6 montre cette hiérarchie pour la fonction de l'exemple 4.8.

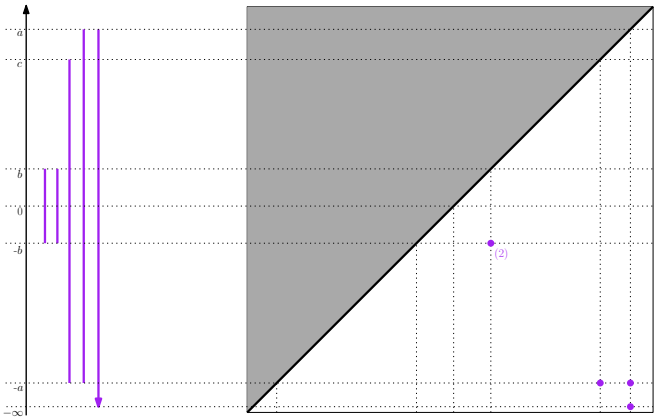


FIGURE 7. Le code-barres issu de la figure 6 (à gauche) et son diagramme de persistance correspondant (à droite). La multiplicité du point $(b, -b)$ dans le diagramme est indiquée entre parenthèses.

Diagrammes de persistance et stabilité. Une autre manière de représenter graphiquement les codes-barres est sous la forme de multi-ensembles de points dans le plan étendu $\mathbb{R} \times [-\infty, +\infty[$, appelés *diagrammes de persistance*, dans lesquels chaque copie du point (a, b) correspond à une copie de l'intervalle $]b, a]$ dans le code-barres associé, avec $a > b \geqslant -\infty$. Voir la figure 7 pour une illustration. Cette représentation alternative contient la même information mais elle offre l'avantage de montrer les codes-barres comme des nuages de points, ou plutôt comme des mesures empiriques, plus facilement interprétables. Par ailleurs, la métrique naturelle entre codes-barres, appelée *distance du goulot de bouteille* (*bottleneck distance* en anglais), peut être interprétée comme une distance de transport partiel entre diagrammes de persistance, eux-mêmes vus

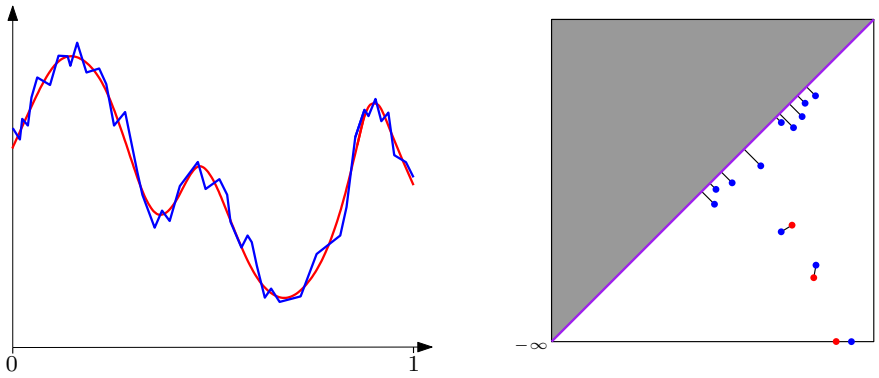


FIGURE 8. Deux fonctions $[0,1] \to \mathbb{R}$ (à gauche) et leurs diagrammes de persistance (à droite) avec, marqué par des segments en noir, l'appariement optimal entre les points des deux diagrammes pour la distance de transport. Dans cet exemple, la distance de transport est légèrement inférieure à la différence des deux fonctions en norme sup.

comme des mesures empiriques. La définition précise de cette distance sera donnée au texte [Car24], mais en attendant nous pouvons déjà la visualiser sur l'exemple de la figure 8. L'interprétation des deux diagrammes dans l'exemple est que chacune des deux fonctions considérées possède trois pics proéminents, le reste des pics (dont les points correspondants dans les diagrammes sont localisés près de la diagonale $y = x$) pouvant être considéré comme du bruit. Le plan de transport optimal associé à la distance du goulot de bouteille entre les diagrammes donne un appariement explicite entre les pics proéminents des deux fonctions, et nous indique par ailleurs d'ignorer les pics non proéminents (en appariant avec la diagonale leurs points correspondants dans les diagrammes).

Ainsi, les diagrammes de persistance fournissent une représentation synthétique des intervalles de persistance des pics d'une fonction, tandis que la distance de transport entre diagrammes indique comment mettre en correspondance au mieux les pics de fonctions différentes selon leurs intervalles de persistance. Cette observation empirique est appuyée par un résultat fondamental de la théorie de la persistance, appelé le *théorème de stabilité*, qui dit en substance que l'opérateur D,

qui associe à toute fonction $X \to \mathbb{R}$ son diagramme de persistance lorsque celui-ci existe, est 1-lipschitzien. Dans notre contexte le résultat s'énonce de la manière suivante – voir encore la figure 8 pour une illustration :

Théorème 4.14. *Pour toutes fonctions $f, g \colon X \to \mathbb{R}$ vérifiant les hypothèses* 4.9 *et* 4.12*, on a l'inégalité suivante, où d_b désigne la distance du goulot de bouteille et $\|\cdot\|_\infty$ désigne la norme sup sur X :*

$$d_b(D(f),\, D(g)) \leqslant \|f - g\|_\infty$$

Dans le texte [Car24] nous énoncerons et démontrerons ce résultat dans le cadre général de la persistance des fonctions réelles, pas seulement celle de leurs pics.

5. Retour à l'application au regroupement

Revenons maintenant à notre application et utilisons la persistance pour corriger les défauts de l'algorithme de clustering présenté à la section 3. Nous n'allons en fait pas modifier l'algorithme en lui-même, mais plutôt ajouter un post-traitement des clusters qu'il produit.

Post-traitement. Étant donné le regroupement $P = \bigsqcup_{\ell=1}^{m} C_\ell$ produit par l'algorithme sur le graphe de voisinage $G = (P, E)$, ainsi que les pics $x_1, \ldots, x_m \in P$ de $\widehat{f}$ dans G associés à chacun des clusters $C_1, \ldots, C_m$, on calcule l'intervalle de persistance de chaque pic x_ℓ à l'intérieur du graphe G, ainsi que son parent (s'il existe) parmi les autres pics. Les détails du calcul importent peu, il suffit de savoir qu'il est possible de le faire en un temps quasi-linéaire en la taille du graphe, par un algorithme similaire à celui de Kruskal pour le calcul d'arbre couvrant minimal [CLRS09, section 23.2]. On obtient ainsi le code-barres de $\widehat{f}$, vue comme une fonction réelle sur le graphe G, vu lui-même comme un espace topologique stratifié par ses sommets et ses arêtes, avec interpolation linéaire des valeurs de $\widehat{f}$ le long des arêtes. Voir la figure 9 pour une illustration.

Étant donné un choix de seuil $\tau \in \mathbb{R}^+$ sur la persistance, on fusionne ensuite itérativement les clusters associés aux pics de persistance plus petite que τ dans les clusters de leurs parents. Notons

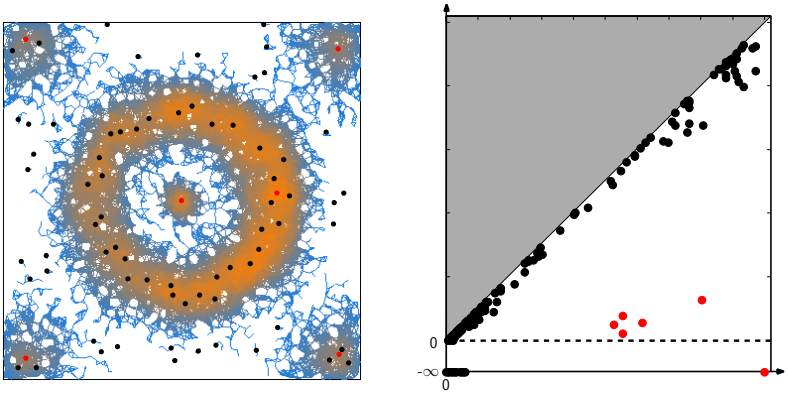


FIGURE 9. À gauche, en rouge et noir : les pics de l'estimateur de densité $\widehat{f}$ dans le graphe de voisinage G de la figure 3. À droite : le diagramme de persistance de $\widehat{f}$, avec, en rouge, six points qui se démarquent clairement du reste du diagramme. Ces points correspondent aux intervalles de persistance des six pics de $\widehat{f}$ en rouge sur la figure de gauche, que l'on peut mettre en correspondance avec les six pics de la vraie densité f (qui sont tout proches, voir la figure 2), le reste des pics de $\widehat{f}$ (en noir) se répartissant dans des zones de faible norme du gradient de f et pouvant être considéré comme du bruit d'après le diagramme de persistance de $\widehat{f}$.

que, dans ce cas particulier, la structure stratifiée de l'espace G et la nature linéaire par morceaux de la fonction $\widehat{f}$ garantissent que tout pic qui n'est pas un maximum global de $\widehat{f}$ sur la composante connexe de G où il se trouve a un parent. Les maxima globaux, quant à eux, ont par définition une persistance infinie. De ce fait, la procédure fusionne bien tous les clusters associés aux pics de persistance plus petite que τ dans d'autres clusters, et *in fine* dans des clusters associés à des pics de persistance plus grande que τ. C'est ainsi que nous avons obtenu le clustering de la figure 1 à partir de celui de la figure 3, par un choix adéquat de seuil τ.

Choix du seuil de persistance. En pratique, pour choisir la valeur du seuil τ on peut s'appuyer sur le diagramme de persistance de $\widehat{f}$ dans G, noté $D(\widehat{f})$, que l'on a calculé lors du post-traitement.

Grâce au théorème de stabilité 4.14, on peut montrer que, sous des hypothèses d'échantillonnage adéquates, le diagramme $D(\widehat{f})$ exhibe une séparation claire entre, d'une part, les intervalles de persistance des pics de $\widehat{f}$ correspondant aux pics de la densité sous-jacente f, et d'autre part, les intervalles de persistance du reste des pics de $\widehat{f}$, qui correspondent à du bruit – voir encore la figure 9 pour une illustration. Grâce à cette séparation, l'utilisateur (ou une méthode statistique) peut sélectionner un seuil τ adapté. Les détails de l'analyse théorique et des garanties associées se trouvent dans l'article [CGOS13] qui a introduit cette approche, appelée *ToMATo* pour *Topological Mode Analysis Tool*.

Références

[Car24] M. Carrière – « Théorie de la persistance (2/2) : stabilité », in *Analyse topologique de données*, Journées X-UPS, Les Éditions de l'École polytechnique, Palaiseau, 2024, ce volume.

[CGOS13] F. Chazal, L. J. Guibas, S. Oudot & P. Skraba – « Persistence-based clustering in Riemannian manifolds », *J. ACM* **60** (2013), no. 6, article no. 41 (38 pages).

[CM02] D. Comaniciu & P. Meer – « Mean shift : A robust approach toward feature space analysis », *IEEE Trans. on Pattern Analysis and Machine Intelligence* **24** (2002), no. 5, p. 603–619.

[CLRS09] T. H. Cormen, C. E. Leiserson, R. L. Rivest & C. Stein – *Introduction to algorithms*, 3e éd., MIT Press, Cambridge, MA, 2009.

[KNF76] W. L. G. Koontz, P. M. Narendra & K. Fukunaga – « A graph-theoretical approach to non-parametric cluster analysis », *IEEE Trans. Comput.* **C-25** (1976), p. 936–944.

[Mil63] J. W. Milnor – *Morse theory*, Annals of Math. Studies, no. 51, Princeton University Press, Princeton, NJ, 1963.

[Oud24] S. Oudot – « Théorie de la persistance (1/2) : structure », in *Analyse topologique de données*, Journées X-UPS, Les Éditions de l'École polytechnique, Palaiseau, 2024, ce volume.

Steve Oudot, Inria Saclay - Ile-de-France & École polytechnique, Bât. Alan Turing, 1 rue Honoré d'Estienne d'Orves, 91120 Palaiseau
E-mail : steve.oudot@inria.fr
Url : https://geometrica.saclay.inria.fr/team/Steve.Oudot/

Journées mathématiques X-UPS
2024, p. 21–35 doi: 10.5802/xups.2024-02

INTRODUCTION RAPIDE À L'HOMOLOGIE

par

Vincent Humilière

Résumé. La théorie de l'homologie associe à tout espace topologique des groupes, de telle sorte que si deux espaces sont homéomorphes alors les groupes associés sont isomorphes. C'est un outil central de topologie dont l'introduction remonte à Poincaré et dont les applications sont innombrables. Ils jouent aussi un rôle clé en analyse topologique des données. Dans ce texte, nous verrons ce que sont ces groupes et ce qu'ils nous disent sur les espaces étudiés.

Table des matières

Dans ce texte nous introduisons un outil fondamental pour l'étude des espaces topologiques, *l'homologie*. Il jouera un rôle central dans les textes suivants.

L'introduction de l'homologie remonte au moins aux travaux de [Poi95], même si des embryons de cette théorie existaient antérieurement (chez Euler par exemple). Le principe est d'associer à n'importe quel espace topologique X une famille d'espaces vectoriels $H_k(X)$ avec $k = 0, 1, 2, \ldots$, de telle sorte que si X et Y sont homéomorphes, alors les espaces vectoriels $H_k(X)$ et $H_k(Y)$ sont isomorphes pour tout k. Les $H_k(X)$ seront appelés *groupes d'homologie* de X. Nous verrons qu'intuitivement, la dimension de $H_k(X)$ groupes d'homologie peut s'interpréter comme un « nombre de cavités de dimension k » renfermées par X.

On peut motiver cette construction comme suit. Le but premier de la topologie est la classification des espaces topologiques à homéomorphisme près. Malheureusement, c'est un problème beaucoup trop difficile, mais en étudiant les groupes d'homologie, ce qui est beaucoup plus simple puisque il s'agit d'algèbre linéaire, on obtient tout de même des informations utiles. Par exemple, pour démontrer que deux espaces ne sont pas homéomorphes, il suffit de montrer qu'ils ont des espaces d'homologies de dimensions différentes !

Les références traitant d'homologie sont innombrables. Par exemple, le livre de [Hat02] est très souvent cité, et notre présentation le suit en partie. Nous conseillons aussi le magnifique site web Analysis Situs du collectif Henri-Paul de Saint Gervais (`https://analysis-situs.math.cnrs.fr/`).

Dans ce texte, on fixe un corps $\mathbb{K}$ qui sera le corps des coefficients de nos groupes d'homologie.

1. Homologie simpliciale

Nous introduisons dans cette partie l'homologie dans le cadre des espaces admettant une structure de *complexes simpliciaux*. Il s'agit d'espaces topologiques construits par induction, ils portent donc une structure combinatoire qui les rend assez simples à étudier. De manière à construire l'intuition, nous commençons par décrire informellement le cas des graphes, puis une généralisation que l'on appellera « graphes bouchés par des triangles ». Si l'on est pressé, on peut passer directement à la partie 1.3 où se trouve la définition formelle de l'homologie simpliciale.

1.1. Cycles dans les graphes. Soit Γ un graphe ayant un ensemble de sommets V et un ensemble d'arêtes E. On suppose que l'ensemble des sommets est ordonné et on oriente chaque arête e de telle sorte qu'elle aille toujours vers un sommet plus grand que son sommet d'origine. On notera les extrémités source(e) et but(e).

On cherche à associer à un graphe des nombres qui soient invariants par déformation, c'est-à-dire par les opérations consistant à écraser une arête pour n'en faire qu'un sommet (figure 1) ou les opérations inverses. Un tel invariant est par exemple le nombre de composantes connexes. Une autre possibilité est le « nombre de boucles » dans le graphe. Le « nombre de boucles » est un terme un peu ambigu, mais on peut en donner une définition précise comme suit.

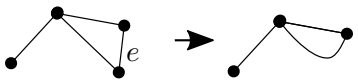


FIGURE 1. Déformation où l'arête e est écrasée sur un sommet.

Appelons *chaîne de sommets* tout ensemble fini $\{(c_i, a_i)\}_{i=1,\dots,\ell}$ de couples (c_i, a_i), où c_i est un sommet de Γ et $a_i \in \mathbb{K}$ un coefficient. On notera une telle chaîne

$$\sum_{i=1}^{\ell} a_i c_i.$$

Définissons de même les *chaînes d'arêtes* comme les ensembles finis de couples constitués d'une arête e_j et d'un coefficient b_j, que l'on notera de manière similaire $\sum e_j b_j$.

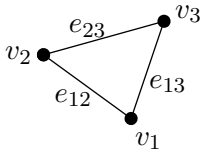
Les chaînes de sommets et d'arêtes forment pour les lois évidentes des $\mathbb{K}$-espaces vectoriels que l'on notera respectivement C_V et C_E. Ces espaces admettent des bases canoniques qui ne sont autres que V et E.

Soit ∂_1 l'unique application linéaire $C_E \to C_V$ définie pour toute arête $e \in E$ par

$$\partial_1(e) = \text{but}(e) - \text{source}(e). \tag{1.1}$$

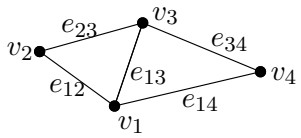
Exemple 1.1. Dans le graphe triangulaire ci-dessous, on a

$$\partial_1(e_{12}+e_{23}) = v_3 - v_2 + v_2 - v_1 = v_3 - v_1.$$



Définition 1.2. On dira qu'une chaîne d'arête c est un *cycle* si $\partial_1 c = 0$. L'ensemble des cycles est un sous-espace vectoriel de C_E noté $H_1(\Gamma)$, c'est le *premier groupe d'homologie* de Γ.

Exemple 1.3. Dans l'exemple 1.1, $e_{12}+e_{23}-e_{13}$ est un cycle. En fait, on peut se convaincre que $H_1(\Gamma)$ est l'ensemble des chaînes de la forme $\lambda(e_{12}+e_{23}-e_{13})$, avec $\lambda \in \mathbb{K}$. Il est donc de dimension 1. Dans l'exemple ci-dessous, on peut se convaincre que les cycles sont exactement les chaînes combinaisons linéaires des chaînes $e_{12}+e_{23}-e_{13}$ et $e_{13}+e_{34}-e_{14}$. L'espace $H_1(\Gamma)$ est donc de dimension 2.



On observe sur les deux graphes ci-dessus que la dimension de $H_1(\Gamma)$ correspond à ce que l'on voudrait appeler le « nombre de boucles dans le graphe ». C'est un fait général.

Par la construction qui précède, nous avons réalisé le « nombre de boucles » comme la dimension d'un espace vectoriel. Le nombre de composantes connexes admet également une telle description. Pour cela, appelons *bord* toute chaîne de sommets qui est l'image par ∂_1 d'une chaîne d'arêtes. Notons que les bords forment un sous-espace vectoriel.

Par exemple, si deux sommets v, v' sont reliés par une arête e, alors la chaîne $v - v' = \partial_1 e$ est un bord. De même, si v'' est un troisième sommet qui est relié à v' par une arête, alors $v''-v = (v''-v')+(v'-v)$

est aussi un bord. Plus généralement, la différence entre les extrémités d'un chemin continu d'arêtes forme toujours un bord. On en déduit que : *deux sommets sont dans la même composante connexe du graphe si et seulement si leur différence est un bord.*

Par conséquent, l'ensemble des composantes connexes s'identifie avec l'espace vectoriel quotient

$$H_0(\Gamma) = C_V/\{\text{bords}\} = C_V/\text{im}(\partial_1). \tag{1.2}$$

Le nombre de composantes connexes correspond alors à la dimension de cet espace.

1.2. Cycles et bords dans les graphes « bouchés par des triangles ». Généralisons légèrement en ne considérant plus seulement des graphes mais des espaces que nous appellerons ici des « graphes bouchés par des triangles » (terminologie absolument pas standard et inventée pour les besoins de cet exposé !), c'est-à-dire des espaces Γ obtenus à partir d'un graphe de sommets V, d'arêtes E, auquel on recolle des triangles de telle sorte que chaque côté soit l'une des arêtes de E. On note T l'ensemble des triangles de Γ (cette définition est volontairement informelle, nous la rendrons plus précise lorsque nous définirons les complexes simpliciaux généraux dans la partie 1.3). Les triangles sont orientés de manière compatible avec l'ordre des sommets.

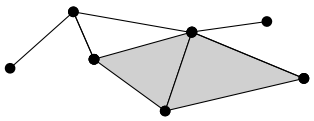


FIGURE 2. Un graphe « bouché par des triangles ». Nous avons ici deux triangles, ce sont les parties grisées.

Comme précédemment, on peut définir des chaînes de sommets, des chaînes d'arêtes et maintenant aussi des chaînes de triangles ; on note C_T l'espace vectoriel qu'elles forment. On peut aussi définir l'opérateur ∂_1 comme en 1.1. On voudrait ici aussi donner un sens au « nombre de boucles » comme précédemment. Pour cela, on voudrait qu'un cycle qui suit le bord d'un triangle ne soit pas compté comme

une boucle. Par exemple, on voudrait que les cycles qui entourent les parties grisées de la figure 2 ne soit pas comptés comme des boucles, de manière à obtenir un nombre invariant par déformation (notamment par l'opération consistant à écraser un triangle sur une arête). Pour cela, on définit l'application linéaire $\partial_2 : C_T \to C_E$ qui à un triangle t dont le bord est constitué d'arêtes e, e', e'' comme sur la figure 3, associe

$$\partial_2(t) = e + e' - e''. \tag{1.3}$$

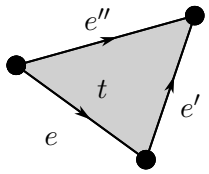


FIGURE 3. Le triangle t et ses arêtes orientées.

On dira qu'une chaîne d'arêtes est un *bord* si elle appartient à l'image de ∂_2. Remarquez que tout bord est un cycle, autrement dit $\partial_2 \circ \partial_1 = 0$.

Définissons alors l'espace vectoriel quotient

$$H_1(\Gamma) = \{\text{cycles}\}/\{\text{bords}\} = \ker(\partial_1)/\mathrm{im}(\partial_2). \tag{1.4}$$

En passant ainsi au quotient, les cycles qui sont des bords sont en quelque sorte « oubliés ».

On peut vérifier dans l'exemple de la figure 2 que, $\ker(\partial_1)$ est de dimension 3, $\mathrm{im}(\partial_2)$ est de dimension 2 (cela correspond aux deux triangles grisés) et donc $\dim H_1(\Gamma) = \dim \ker(\partial_1) - \dim \mathrm{im}(\partial_1)$ est de dimension 1. Cela correspond bien au nombre de boucles que l'on souhaite compter. Ici, il y a une seule boucle qui n'est pas bouchée par un triangle.

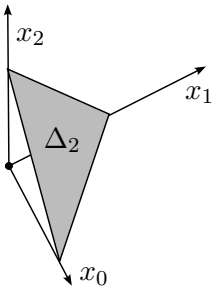
En plus du nombre de composantes connexes et du nombre de boucles, les graphes bouchés par des triangles ont un troisième invariant intéressant, le nombre de cavités.

Par exemple, si Γ est obtenu en prenant tous les sommets, toutes les arêtes et toutes les faces d'un tétraèdre, alors on voit qu'il renferme

une cavité. Si l'on fait de même à partir de deux tétraèdres collés le long d'une face, il y aura deux cavités si l'on garde la face de recollement, mais une seule si on l'enlève.

On peut formaliser ce compte de cavités suivant le même procédé que précédemment en disant qu'une chaîne de triangles c est un *cycle* si $\partial_2(c) = 0$, et en définissant $H_2(\Gamma)$ comme l'ensemble des cycles. Le nombre de cavités du graphe bouché correspond alors à la dimension de $H_2(\Gamma)$.

La construction des graphes bouchés peut être amplement généralisée, en bouchant des cavités par des tétraèdres, puis en réitérant le procédé en dimension supérieure. On arrive alors à la définition de complexe simplicial.



1.3. Complexes simpliciaux. Pour tout entier naturel n, on appelle n-simplexe standard l'ensemble

$$\Delta_n = \left\{t \in (\mathbb{R}_{\geqslant 0})^{n+1} : \textstyle\sum_{i=0}^{n} t_i = 1\right\}.$$

Les éléments $(v_0, \ldots, v_n)$ (ordonnés dans cet ordre) de la base canonique de $\mathbb{R}^{n+1}$ sont les *sommets* de Δ_n. Si $n \geqslant 1$, les *faces* de Δ_n sont les

$$F_i = \Delta_n \cap \{t_i = 0\}, \quad i = 0, \ldots, n.$$

Chacune d'entre elle s'identifie naturellement à Δ_{n-1}. L'*intérieur* $\mathring{\Delta}_n$ de Δ_n est l'ensemble Δ_n privé de ses faces.

Définition 1.4. Une structure de complexe simplicial[(1)] sur un espace topologique X est une famille d'applications continues $(\sigma_\alpha)_{\alpha \in A}$, où chaque σ_α est une application $\Delta_n \to X$ pour un entier n dépendant de α, vérifiant :

[(1)]La terminologie correcte est « structure de Δ-complexe », cf. [Hat02].

(1) σ_α est injective en restriction à l'intérieur $\mathring{\Delta}_n$ de Δ_n,
(2) tout point de X est dans l'image d'exactement l'une des $\sigma_\alpha|_{\mathring{\Delta}_n}$,
(3) toute restriction d'une application σ_α à une face de Δ_n est l'une des $\sigma_\beta : \Delta_{n-1} \to X$,
(4) une partie U de X est ouverte si et seulement si $\sigma_\alpha^{-1}(U)$ est ouverte pour tout α.

Les applications σ_α sont appelées *n-simplexes* de la structure simpliciale. L'entier n maximal de la famille des σ_α est appelé la *dimension* du complexe simplicial. Le cas de la dimension 1 est celui des graphes, celui de la dimension 2 correspond au cas des graphes bouchés par des triangles de la section précédente. La définition ci-dessous nous sera utile dans le texte [Oud24].

Définition 1.5. Un sous-complexe simplicial d'un complexe simplicial X est un sous-espace topologique de X qui admet une structure de complexe simplicial dont tous les simplexes sont des simplexes de X.

Dans la suite, on suppose donnée une structure de complexe simplicial sur un espace X comme dans la définition. Les notions de chaînes, cycles, bords, se généralisent comme suit.

Définition 1.6. On appelle *n-chaîne* tout ensemble fini de couples $(\sigma_1, a_1), \dots, (\sigma_\ell, a_\ell)$ où chaque σ_i est l'un des $\sigma_\alpha : \Delta_n \to X$ et $a_i \in \mathbb{K}$. On note une telle n-chaîne sous la forme

$$\sum_{i=1}^{\ell} a_i \sigma_i.$$

Les n-chaînes forment pour les lois évidentes un espace vectoriel que l'on notera C_n ·

Le n-ième opérateur de bord est l'application linéaire $\partial_n : C_n \to C_{n-1}$ telle que pour tout n-simplexe σ_α,

$$\partial_n(\sigma_\alpha) = \sum_{i=0}^{n} (-1)^i \sigma_\alpha|_{F_i}. \tag{1.5}$$

Un *n-cycle* est une n-chaîne c telle que $\partial_n c = 0$. Un *n-bord* est une n-chaîne c qui est dans l'image de ∂_{n+1}.

Remarquez que la formule (1.5) généralise directement les formules (1.1) et (1.3). Par convention, on pose $\partial_0 = 0$, autrement dit, toutes les 0-chaînes sont des cycles.

Exercice 1.7. Vérifier qu'un bord est toujours un cycle, autrement dit $\partial_n \circ \partial_{n+1} = 0$.

Intuitivement, les n-cycles correspondent à des « cavités de dimension n » et les n-bords correspondent aux cavités qui sont « remplies ». Nous pouvons enfin définir l'homologie simpliciale.

Définition 1.8. Pour tout entier $n \geqslant 0$, le *n-ième groupe d'homologie simpliciale* de X est l'espace vectoriel quotient

$$H_k^{\text{simpl}}(X) = \{\text{cycles}\}/\{\text{bords}\} = \ker(\partial_n)/\text{im}(\partial_{n+1}).$$

Notez que cette définition généralise à la fois (1.2) et (1.4). Nous verrons plus bas que cet espace vectoriel ne dépend pas (à isomorphisme près) du choix de la structure simpliciale sur X, ce qui justifie la notation. Nous allons maintenant voir quelques exemples de calculs simples.

1.4. Calcul de l'homologie simpliciale.

Exemple 1.9 **(Homologie en degré 0).** L'argument développé à la fin de la partie 1.1 montre que la dimension de $H_0^{\text{simpl}}(X)$ est toujours égale au nombre de composantes connexes de X.

Exemple 1.10 **(Sphères).** La sphère $\mathbb{S}^2 = \{(x, y, z) \in \mathbb{R}^3\}$ admet une structure simpliciale obtenue en traçant un tétraèdre sur sa surface (voir figure 4).

Cette structure a quatre sommets, six arêtes et quatre triangles. Comme $\mathbb{S}^2$ a une unique composante connexe, $H_0^{\text{simpl}}(\mathbb{S}^2)$ est de dimension 1. Par (1.2), ∂_1 est donc de rang 3 et par le théorème du rang, $\ker \partial_1$ est de dimension 3. On peut ensuite se convaincre que la somme des quatre triangles est un cycle, et que tous les autres 2-cycles sont colinéaires à celui-ci. Autrement dit, $\ker \partial_2$ est de dimension 1. Par le théorème du rang, son image est de dimension 3. Comme $\text{im}\, \partial_2$ est inclus dans $\ker \partial_1$ et qu'ils ont la même dimension, ils coïncident.

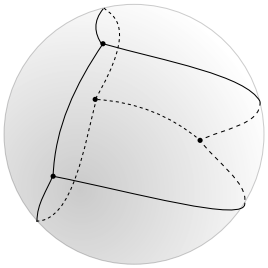


FIGURE 4. Un tétraèdre tracé sur une sphère.

En conclusion, on obtient

$$H_k^{\text{simpl}}(\mathbb{S}^2) \simeq \begin{cases} \mathbb{K} & \text{si } k = 0, 2, \\ \{0\} & \text{sinon.} \end{cases}$$

De la même manière, on peut donner une structure simpliciale à toute sphère $\mathbb{S}^n$ en l'identifiant au bord d'un $n+1$-simplexe et calculer les groupes d'homologie correspondant. On obtient

Proposition 1.11.

$$H_k^{\text{simpl}}(\mathbb{S}^n) \simeq \begin{cases} \mathbb{K} & \textit{si } k = 0, n, \\ \{0\} & \textit{sinon.} \end{cases}$$

Ce résultat est compatible avec l'intuition que l'homologie compte les cavités : une sphère de dimension n a bien une seule cavité et celle-ci est bien est de dimension n.

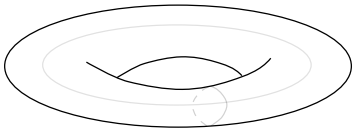


FIGURE 5. Le tore $\mathbb{T}^2$

On peut aussi regarder l'exemple du tore $\mathbb{T}^2$ (figure 5). Celui-ci peut être obtenu en recollant les côtés opposés d'un carré (figure 6). Cette description du tore donne une structure simpliciale dont les trois arêtes sont les côtés du carré (identifiés par paires) et une diagonale du carré. Cette structure a un sommet, trois arêtes et deux triangles.

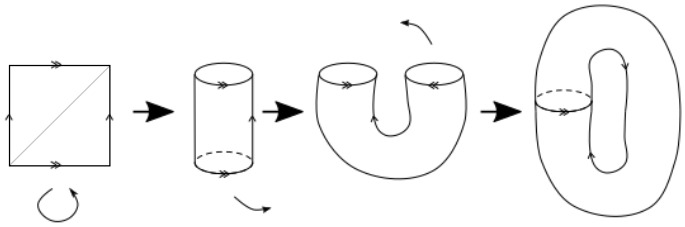


FIGURE 6. Le tore $\mathbb{T}^2$ à partir d'un carré

Exercice 1.12. À partir de la structure simpliciale ci-dessus, vérifier que

$$H_k^{\text{simpl}}(\mathbb{T}^2) \simeq \begin{cases} \mathbb{K} & \text{si } k = 0, 2, \\ \mathbb{K}^2 & \text{si } k = 1. \end{cases}$$

En déduire que le tore $\mathbb{T}^2$ n'est pas homéomorphe à la sphère $\mathbb{S}^2$.

Les générateurs du groupe d'homologie de degré 1 viennent des deux arêtes qui correspondent aux côtés du carré. Ils sont représentés sur la figure 5 par les courbes grises.

Terminons par une remarque générale sur le calcul de l'homologie simpliciale.

Remarque 1.13. Si l'on dispose d'une structure simpliciale explicite, l'homologie simpliciale, ou au moins sa dimension, est très facile à calculer. En effet, la dimension de l'homologie est donnée par

$$\dim H_k^{\text{simpl}}(X) = \dim \ker \partial_k - \dim \operatorname{im} \partial_{k+1}.$$

Donc par le théorème du rang,

$$\dim H_k^{\text{simpl}}(X) = \dim C_k(X) - \operatorname{rang}(\partial_k) - \operatorname{rang}(\partial_{k+1}), \tag{1.6}$$

où $\dim C_k(X)$ n'est autre que le nombre de k-simplexes dans la décomposition simpliciale de X.

Il s'agit donc simplement de calculer les rangs des applications ∂_n. Or ces applications sont données par leur matrice dans les bases canoniques des différents $C_n(X)$ (constituées des σ_α). Ces calculs peuvent donc être réalisés efficacement par l'algorithme du pivot de Gauss (en temps $O(p^2 q)$ pour une matrice de dimension $p \times q$).

2. Homologie singulière

L'homologie simpliciale est facile à calculer, mais elle présente quelques inconvénients majeurs, à commencer par le fait qu'expliciter une structure de complexe simplicial n'est pas toujours chose aisée. La théorie de l'homologie singulière que nous allons maintenant présenter est moins facile à calculer a priori, mais plus pratique et flexible.

2.1. Définition formelle. On fixe dans cette partie un espace topologique quelconque X. Dans la définition de l'homologie simpliciale, on appelait n-chaînes les combinaisons linéaires formelles formées à partir des n-simplexes σ_α de la décomposition simpliciale. Ici, nous allons considérer les combinaisons linéaires de *tous* les n-simplexes ; c'est ce qui distingue la définition ci-dessous de la définition 1.6.

Définition 2.1. Un *n-simplexe* est une application continue $\Delta_n \to X$. Une *n-chaîne singulière* est un ensemble fini de couples

$$(\sigma_1, a_1), \dots, (\sigma_\ell, a_\ell),$$

où chaque σ_i est un n-simplexe et $a_i \in \mathbb{K}$. On la note $\sum_{i=1}^{\ell} a_i \sigma_i$. Les n-chaînes singulières forment pour les lois évidentes un espace vectoriel que l'on notera $C_n(X)$.

Le n-ième opérateur de bord singulier est l'application linéaire $\partial_n : C_n(X) \to C_{n-1}(X)$ telle que pour tout n-simplexe σ,

$$\partial_n(\sigma) = \sum_{i=0}^{n} (-1)^i \sigma|_{F_i}. \tag{2.1}$$

Comme plus haut, on appellera *n-cycles* les éléments de $\ker \partial_n$ et *n-bords* ceux de $\operatorname{im} \partial_{n+1}$, et nous avons l'analogue de la définition 1.8.

Définition 2.2. Pour tout entier $n \geqslant 0$, le *n-ième groupe d'homologie singulière* de X est l'espace vectoriel quotient

$$H_n^{\mathrm{sing}}(X) = \{\text{cycles}\}/\{\text{bords}\} = \ker(\partial_n)/\operatorname{im}(\partial_{n+1}).$$

Remarquez que cette construction s'applique à n'importe quel espace topologique X. Elle ne dépend d'aucun autre choix. La construction passe par les espaces $C_n(X)$ qui sont (presque) toujours de

dimension infinie et il est difficile de se faire une intuition a priori sur ce que calcule cette homologie. Cependant, le résultat suivant montre qu'en présence d'une décomposition simpliciale, les deux homologies coïncident.

Théorème 2.3. *Soit X un espace topologique muni d'une structure de complexe simplicial. Alors, pour tout entier n, on a un isomorphisme d'espaces vectoriels*

$$H_n^{\mathrm{sing}}(X) \simeq H_n^{\mathrm{simpl}}(X).$$

Nous ne démontrerons pas cet énoncé ici. Il implique que l'homologie simpliciale ne dépend pas du choix de structure simpliciale. Ce théorème et les calculs de la partie 1.4 donnent des exemples de calcul de l'homologie singulière. Il montre aussi que l'homologie singulière peut aussi être pensée intuitivement comme le nombre de cavités renfermées par l'espace X. On notera dorénavant toutes les homologies par $H_n(X)$, ou $H_*(X)$ si l'on souhaite considérer tous les n possibles.

2.2. Action des fonctions continues en homologie et invariance par déformation. Un intérêt très important de la définition « singulière » de l'homologie est le fait qu'une application continue entre espaces topologiques induit une application linéaire entre groupes d'homologie de manière directe.

Soit $f : X \to Y$ une application continue entre espaces topologiques. À tout n-simplexe singulier $\sigma : \Delta_n \to X$, on peut associer le n-simplexe singulier $f \circ \sigma : \Delta_n \to Y$, ce qui induit des applications linéaires entre chaînes

$$f_* : C_n(X) \longrightarrow C_n(Y)$$

pour tout entier $n \geqslant 0$.

Remarquez que f_* est compatible avec les opérateurs bords. En effet, la formule (2.1) donne directement

$$f_* \circ \partial_n^X = \partial_n^Y \circ f_*,$$

où ∂_n^X et ∂_n^Y sont les opérateurs bords définis respectivement sur X et Y. On en déduit que f_* envoie cycle sur cycle et bord sur bord :

$$f_*(\ker \partial_n^X) \subset \ker \partial_n^Y \quad f_*(\mathrm{im}\ \partial_{n+1}^X) \subset \mathrm{im}\ \partial_{n+1}^Y.$$

Ceci implique donc que f_* induit des applications linéaires, encore notées f_*,

$$f_* : H_*(X) \longrightarrow H_*(Y).$$

Soit maintenant $g : Y \to Z$ une autre application continue. Pour tout simplexe σ, l'associativité de la composition donne $(g \circ f) \circ \sigma = g \circ (f \circ \sigma)$ et l'on en déduit $(g \circ f)_* = g_* \circ f_* : C_n(X) \to C_n(Z)$. Par ailleurs, on vérifie facilement que l'identité agit comme l'identité sur les chaînes. Nous obtenons donc :

Proposition 2.4. *Pour toutes applications continues $f : X \to Y$ et $g : Y \to Z$ on a*

$$(g \circ f)_* = g_* \circ f_* : H_*(X) \longrightarrow H_*(Z).$$

De plus, $(\mathrm{id}_X)_* = \mathrm{id}_{H_*(X)}$.

On en déduit immédiatement :

Corollaire 2.5. *Si f est un homéomorphisme alors f_* est un isomorphisme d'espaces vectoriels.*

Remarque 2.6. Les propriétés ci-dessus s'expriment dans le langage de la théorie des catégories en disant simplement que l'homologie est un « foncteur de la catégorie des espaces topologiques vers celle des espaces vectoriels ».

Les applications induites en homologie f_* ont une propriété très importante qui est leur invariance par déformation. Notons $\mathcal{C}^0(X, Y)$ l'espace des fonctions continues de X dans Y, muni de la topologie compacte-ouverte (si Y est un espace métrique, il s'agit de la topologie de la convergence uniforme sur tout compact).

Proposition 2.7. *Soient $f_0, f_1 : X \to Y$ deux fonctions continues* homotopes*, c'est-à-dire reliées par un chemin continu dans $\mathcal{C}^0(X, Y)$. Alors,*

$$(f_0)_* = (f_1)_* \quad : H_*(X) \longrightarrow H_*(Y).$$

Deux espaces obtenus l'un de l'autre par une « déformation continue » auront donc la même homologie. Voici un exemple d'utilisation.

Exemple 2.8. Considérons les applications $p : \mathbb{R}^{n+1} \smallsetminus \{0\} \to \mathbb{S}^n$ et $i : \mathbb{S}^n \to \mathbb{R}^{n+1} \smallsetminus \{0\}$ définies par

$$p(x) = x/\|x\|, \quad i(x) = x.$$

Clairement, $p \circ i = \mathrm{id}_{\mathbb{S}^n}$, donc $p_* \circ i_* = \mathrm{id}_{H_*(\mathbb{S}^n)}$. Par ailleurs, $i \circ p$ peut être déformé sur l'identité de $\mathbb{R}^{n+1} \smallsetminus \{0\}$ via le chemin d'applications continues $f_t(x) = (1-t)x + tx/\|x\|$ qui vérifie $f_0 = \mathrm{id}$ et $f_1 = i \circ p$. D'où l'identité $i_* \circ p_* = \mathrm{id}_{H_*(\mathbb{R}^{n+1} \smallsetminus \{0\})}$. On conclut que i_* et p_* sont inverses l'une de l'autre et donc pour $k \geqslant 0$,

$$H_k(\mathbb{R}^{n+1} \smallsetminus \{0\}) \simeq H_k(\mathbb{S}^n) \simeq \begin{cases} \mathbb{K} & \text{si } k = 0, n, \\ \{0\} & \text{sinon.} \end{cases}$$

2.3. Le théorème d'invariance du domaine. Nous ne pouvons pas conclure ce texte sans citer cette application très classique de l'homologie.

***Théorème 2.9* (Brouwer, 1912).** *Soient n, m des entiers naturels tels que $\mathbb{R}^n$ est homéomorphe à $\mathbb{R}^m$. Alors $n = m$.*

Démonstration. Le cas où $m = 0$ ou $n = 0$ étant évidents, on suppose que $n, m > 0$. Soit $f : \mathbb{R}^n \to \mathbb{R}^m$ un homéomorphisme. Quitte à composer f par une translation, on peut supposer que $f(0) = 0$ et donc que f se restreint à un homéomorphisme de $\mathbb{R}^n \smallsetminus \{0\}$ vers $\mathbb{R}^m \smallsetminus \{0\}$. Pour des raisons de connexité, si $n = 1$, alors $m = 1$ aussi. Supposons donc dorénavant que $n > 1$. D'après l'exemple 2.8, $H_{n-1}(\mathbb{R}^n \smallsetminus \{0\})$ est non trivial. D'après le corollaire 2.5, l'homologie $H_{n-1}(\mathbb{R}^m \smallsetminus \{0\})$ doit donc aussi être non triviale. On conclut avec l'exemple 2.8 que $n = m$. □

Références

[Hat02] A. Hatcher – *Algebraic topology*, Cambridge University Press, Cambridge, 2002.

[Oud24] S. Oudot – « Théorie de la persistance (1/2) : structure », in *Analyse topologique de données*, Journées X-UPS, Les Éditions de l'École polytechnique, Palaiseau, 2024, ce volume.

[Poi95] H. Poincaré – « Analysis situs », *J. École Polytech. (2)* **1** (1895), p. 1–123.

Vincent Humilière, Sorbonne Université and Université de Paris, CNRS, IMJ-PRG, F-75006 Paris, France, & Institut Universitaire de France
E-mail : vincent.humiliere@imj-prg.fr
Url : https://webusers.imj-prg.fr/~vincent.humiliere/