PRÉFACE

Les outils issus de la topologie ont récemment eu un impact sur l'analyse de données. L'un des développements est l'homologie persistante. Supposons que l'on dispose de données sous la forme d'un nuage de points, c'est-à-dire d'un ensemble de points. Si cet ensemble a été échantillonné à partir d'un objet, on aimerait utiliser ce nuage pour déduire les propriétés de l'objet. L'homologie persistante applique les outils de la topologie algébrique à cette fin. On la retrouve notamment dans de nouvelles classes de descripteurs pour les données, utilisées en apprentissage automatique. Les textes de ces journées forment un tout, en introduisant différentes facettes de ce type de questions.

Dans un texte introductif [Oud24a], *Steve Oudot* donne des éléments de contexte sur l'analyse topologique de données et son développement. Puis, de manière informelle, il présente les idées qui sous-tendent la théorie de la persistance topologique, qui rassemble les fondements mathématiques du domaine.

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. Cette théorie est un outil central de topologie dont l'introduction remonte à Poincaré et dont les applications sont innombrables. Ces groupes jouent aussi un rôle clé en analyse topologique des données, ce qui nous est expliqué par *Vincent Humilière* [Hum24a].

Steve Oudot utilise l'homologie dans le texte [Oud24b] pour introduire la théorie de la persistance topologique, notamment dans ses aspects algébriques.

Dans le texte [Car24a], *Mathieu Carrière* étudie différentes répercussions du théorème de stabilité en analyse de données et en inférence géométrique et statistique. Celui-ci garantit que des diagrammes de persistance issus des sous-niveaux de fonctions proches en norme infinie, sont eux-mêmes proches au sens de la distance « du goulot de bouteille ».

De manière surprenante, les idées issues de l'analyse topologique de données, et la théorie de la persistance en particulier, ont eu des applications très récentes en mathématiques fondamentales. *Vincent Humilière* en explique deux dans le texte [Hum24b].

Enfin, dans le texte [Car24b], *Mathieu Carrière* formalise les bases de l'apprentissage automatique supervisé et non-supervisé, ainsi que les différentes approches permettant l'incorporation des diagrammes de persistance dans les modèles standards via les méthodes à noyaux.

Nous tenons à remercier la direction de l'École polytechnique, la Direction des Services de l'Enseignement et le Centre Poly-Média, pour l'aide matérielle importante qu'ils ont apportée à la préparation de ces journées et à la publication de ce volume. Nos remerciements vont aussi au Labex Mathématique Hadamard pour le financement des captations vidéos des exposés, ainsi qu'à Hélios Azzollini pour leur réalisation remarquable, mises en ligne sur la chaîne Youtube de l'École polytechnique : `https://www.youtube.com/playlist?list=PLrRN3yszYHZkR9vyUeOVkcF6yy4FjgkMn`

Nous remercions enfin le secrétariat du Centre de Mathématiques Laurent Schwartz, notamment Carole Juppin, qui assure chaque année le bon déroulement des journées.

Pascale Harinck, Alain Plagne et Claude Sabbah

PRÉSENTATION GÉNÉRALE

Cet ouvrage est une courte introduction aux fondements mathématiques de l'analyse topologique de données, communément appelée TDA d'après l'anglais *Topological Data Analysis*. Ce domaine de l'intelligence artificielle s'est développé à partir des années 2000 et a connu un fort essor du fait de son positionnement transversal, à l'interface entre l'algèbre, la topologie, la géométrie, l'algorithmique, les statistiques, l'optimisation et l'apprentissage machine. Il combine en effet (et contribue à développer) une grande variété d'outils mathématiques et informatiques, ce qui fait tout son attrait sur le plan scientifique.

Les fondements mathématiques du domaine sont regroupés sous le terme de *théorie de la persistance*. Les idées de base qui la sous-tendent ne sont en soi pas nouvelles et remontent à la théorie de Morse : on regarde les sous-niveaux de fonctions réelles et on utilise un invariant algébrique (l'*homologie*) pour encoder l'évolution de la topologie à travers ces sous-niveaux. Les outils utilisés sont toutefois plus avancés que la théorie de Morse car les fonctions considérées sont à peu près arbitraires, en particulier elles peuvent être non lisses ou même discontinues, et dans les développements les plus récents de la théorie elles peuvent même être à valeurs dans $\mathbb{R}^n$. Les aspects algébriques de la théorie de la persistance utilisent et développent des outils issus de la topologie algébrique, de la théorie des représentations, de la théorie des faisceaux, ou encore de l'algèbre commutative. À chacun de ces domaines la TDA offre de nouvelles perspectives d'application, tout en éclairant certaines des grandes questions du

domaine d'une lumière originale et en proposant des manières inédites de les aborder, notamment à travers le prisme de la stabilité. La théorie de la persistance elle-même trouve des applications au-delà de l'analyse de données, dans d'autres domaines des mathématiques fondamentales comme la topologie symplectique, la géométrie spectrale ou encore l'analyse complexe. La richesse de toutes ces interactions ne peut être vraiment mise en valeur dans une introduction courte au sujet comme celle présentée dans ce livre. Pour cela nous renvoyons le lecteur vers d'autres ouvrages plus avancés et plus complets.

La partie centrale du livre (textes [Oud24b] et [Car24a]) se concentre sur le cadre le plus standard de la théorie de la persistance, celui des fonctions réelles sur des espaces topologiques. Elle fournit une introduction pédestre au sujet, insistant sur les principaux résultats de structure et de stabilité et fournissant juste ce qu'il faut d'arguments de preuve pour convaincre le lecteur de la validité des énoncés. Elle est suivie d'une partie applicative (textes [Hum24b] et [Car24b]) qui présente une sélection d'applications en mathématiques fondamentales d'une part, en analyse de données d'autre part. Le tout est précédé de deux textes introductifs, l'un (texte [Oud24a]) présentant quelques-unes des principales idées de la théorie de manière accessible dans un cadre applicatif particulier, l'autre (texte [Hum24a]) introduisant les bases de l'homologie qui sont utilisées dans la suite. L'ensemble forme un ouvrage court que l'on peut aisément glisser dans son sac et lire de manière linéaire le temps d'une escapade. Son contenu devrait être lisible par les étudiants dès la licence, à condition qu'ils aient un bagage en algèbre linéaire et bilinéaire ainsi qu'en topologie générale. Il devrait également intéresser les mathématiciens d'autres disciplines qui recherchent une introduction brève au sujet.

La théorie de la persistence en analyse de données et au-delà

Dans le contexte de l'analyse de données, la théorie de la persistance est utilisée dans la chaîne de traitement pour engendrer de nouvelles représentations pour les données, comme illustré dans la figure 1. En détails : en partant des données, vues comme un

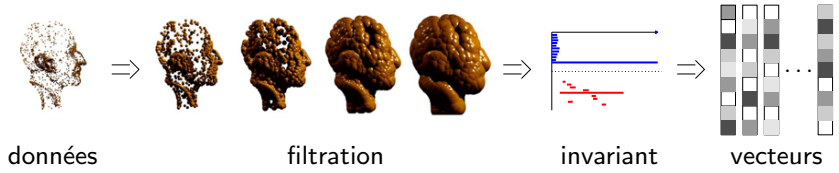

FIGURE 1. La chaîne de traitement de la TDA.

nuage de points dans un espace métrique (l'espace euclidien $\mathbb{R}^3$ dans l'exemple), on construit une famille croissante (pour l'inclusion) d'espaces topologiques, appelée une *filtration*. Pour cela on regarde les sous-niveaux d'une fonction, choisie en fonction du contexte, dans l'exemple la distance aux données dans l'espace ambiant. L'homologie de cette filtration nous donne un objet algébrique appelé *module de persistance*, sur lequel on calcule un ou plusieurs invariants, qui, dans le cadre de cet ouvrage, prennent la forme de code-barres comme illustré dans la figure. Ces codes-barres sont ensuite transformés en vecteurs qui servent de nouvelle représentation pour les données d'entrée et peuvent être intégrés dans d'autres chaînes de traitement comme par exemple des réseaux de neurones. À noter que les données fournies en entrée ne sont pas forcément les données initiales du problème : elles peuvent en être une version déjà transformée, ce qui fait que la TDA peut s'insérer à divers endroits (pas seulement au début) de la chaîne de traitement.

Comme nous le verrons dans le texte [Car24b], la pertinence de l'approche décrite ci-dessus repose sur trois propriétés fondamentales des modules de persistance, détaillées dans les textes [Oud24b] et [Car24a] :

- le fait que l'on puisse les définir à partir des sous-niveaux de n'importe quelle fonction réelle sur un espace topologique quelconque ;
- le fait que, structurellement, ils soient entièrement caractérisés par leur code-barres, et ce, sous des hypothèses très faibles ;
- enfin le fait qu'une métrique canonique puisse être mise sur l'ensemble de ces codes-barres, de manière à les rendre stables par perturbation des fonctions (et donc des données) d'entrée.

Ces trois propriétés rendent également pertinent l'usage des modules de persistance dans d'autres contextes, y compris en mathématiques fondamentales comme il a été dit plus haut. Dans ces contextes, le rôle joué par la stabilité des modules de persistance et de leurs invariants est proéminent et permet d'aborder des questions réputées difficiles sous un angle nouveau. Les exemples présentés dans le texte [Hum24b] donneront une idée plus précise au lecteur.

Quelques perspectives sur le sujet. Le développement de la théorie de la persistance et de ses applications, que ce soit en intelligence artificielle ou en mathématiques fondamentales, est un thème de recherche très actif. Une présentation exhaustive des questions ouvertes qui concentrent actuellement l'attention des chercheurs serait hors de propos dans cette introduction. Toutefois, nous souhaitons mentionner deux de ces problématiques qui comptent parmi les plus importantes du fait de leur impact applicatif majeur.

La première problématique concerne l'étude de la dérivabilité de la chaîne de traitement de la TDA, en particulier de la construction des codes-barres. Nous verrons dans les textes qui suivent que la construction des codes-barres est pour une grande part combinatoire, ce qui rend la définition d'une dérivée difficile voire impossible de prime abord. Et pourtant, grâce aux propriétés de stabilité des codes-barres il est possible de définir une dérivée presque partout. En effet, cette stabilité s'exprime grossièrement de la manière suivante : l'opérateur qui associe son code-barres à une fonction réelle est lipschitzien. De ce fait, par un résultat bien connu de théorie géométrique de la mesure (le théorème de Rademacher), l'opérateur est différentiable presque partout. Reste à définir le bon cadre théorique pour formaliser cet énoncé, trouver des formules explicites pour la différentielle, et étudier ce qu'il se passe au voisinage des points singuliers. C'est l'objet de toute une série de travaux pionniers récents du domaine, qui esquissent une théorie du calcul différentiel et de l'optimisation dans l'espace des codes-barres, et rendent possible leur usage dans de nouveaux contextes comme par exemple en apprentissage profond.

La deuxième problématique concerne l'étude de la topologie des sous-niveaux de fonctions à valeurs dans $\mathbb{R}^n$, appelée communément

multi-persistance car les filtrations et les modules de persistance associés sont à plusieurs paramètres. Cette extension de la théorie s'avère infiniment plus complexe que sa version de base, en lien avec des questions notoirement difficiles et ouvertes issues d'autres domaines des mathématiques comme par exemple celle de la classification des modules indécomposables sur les algèbres de type sauvage en théorie des représentations. Dans ce contexte il n'existe pas de notion canonique de code-barres pour les modules de persistance, et tout l'enjeu est de développer des invariants alternatifs qui soient à la fois calculables, stables, et suffisamment fins pour les applications visées. De nombreuses approches sont actuellement explorées, qui abordent la question sous des angles très divers, comme par exemple ceux de la théorie de Cerf, de la théorie des faisceaux, de l'algèbre homologique, ou encore de la théorie des ordres. Ces multiples développements donnent lieu à un foisonnement de nouvelles propositions d'invariants, dont l'analyse théorique et la validation expérimentale occuperont sans doute la communauté pendant de nombreuses années. L'espoir étant qu'à terme émerge une théorie bien fondée de la multi-persistance, au même titre qu'a émergé une version de base pour les fonctions réelles telle que présentée dans cet ouvrage.

Mathieu Carrière, Vincent Humilière et Steve Oudot

Références

[Car24a] 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.

[Car24b] ———, « Applications en apprentissage automatique », in *Analyse topologique de données*, Journées X-UPS, Les Éditions de l'École polytechnique, Palaiseau, 2024, ce volume.

[Hum24a] V. Humilière – « Introduction rapide à l'homologie », in *Analyse topologique de données*, Journées X-UPS, Les Éditions de l'École polytechnique, Palaiseau, 2024, ce volume.

[Hum24b] ———, « Applications de la théorie de la persistance en géométrie », in *Analyse topologique de données*, Journées X-UPS, Les Éditions de l'École polytechnique, Palaiseau, 2024, ce volume.

[Oud24a] S. Oudot – « Introduction à la théorie de la persistance à travers un exemple d'application », in *Analyse topologique de données*, Journées X-UPS, Les Éditions de l'École polytechnique, Palaiseau, 2024, ce volume.

[Oud24b] ———, « 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.