Barycentres de Wasserstein pour les diagrammes de persistance

Julien Tierny

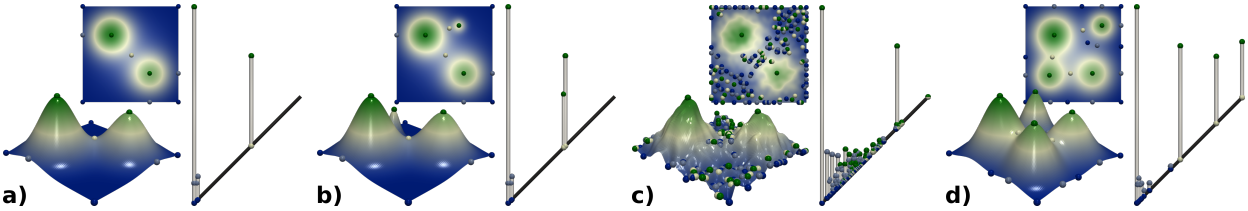

Fig. 1. Le sujet en une image – L'homologie persistante est un outil théorique puissant, qui permet en pratique d'introduire une mesure de bruit sur les structures topologiques, comme les singularités (sphères de couleur) dans cet exemple de carte d'élévation. Cette mesure de bruit, appelée *persistance*, permet de visualiser et de mesurer des structures topologiques à plusieurs échelles d'importance et d'extraire efficacement et avec précision les structures d'intérêt dans un jeu de données. La persistance est souvent associée au *diagramme de persistance* (à droite sur ces exemples), qui donne une représentation visuelle de la distribution des structures topologiques (ici des singularités) en fonction de leur plage de valeur dans les données. Ces diagrammes, grâce à leur stabilité [10], jouent un rôle central en analyse topologique de données: ils constituent en effet une représentation réduite des données particulièrement pertinente, qui capture les structures les plus importantes des données. Ainsi, il est possible dans de nombreuses applications d'effectuer des analyses avancées directement sur le diagramme plutôt que sur les données initiales, qui sont en général plusieurs ordres de grandeur plus volumineuses. Dans de nombreuses applications, il est nécessaire de regrouper par similarité des jeux de données (*clustering*). Dans ce cadre, il serait très efficace en temps d'effectuer ce clustering sur les diagrammes de persistance plutôt que sur les données initiales. La plupart des algorithmes de clustering ont recours au calcul de barycentres entre jeux de données (c'est le cas du *k-means*). Cependant, il n'existe pas de manière établie de calculer un barycentre entre plusieurs diagrammes de persistance (de calculer un diagramme de persistance *moyen*). Dans ce stage, nous souhaitons répondre à ce problème en définissant des barycentres de diagrammes de persistance, afin de re-grouper entre eux des diagrammes similaires, soit dans l'exemple ci-dessus, de regrouper les diagrammes A, B et C dans une même classe (deux gaussiennes) et d'affecter le diagramme D à une seconde classe (quatre gaussiennes).

1 Contexte

L'analyse topologique de données (TDA) [8, 13, 29, 34, 35] est une discipline à cheval entre informatique et mathématiques appliquées qui propose d'analyser des données complexes au vu de leur structure, de leur topologie [28]. Elle connait un essor important depuis quelques années, dû principalement à ses succès récents en *data science* [29, 35] et en *machine learning* [7].

Parmi les différents outils d'analyse développés en TDA (comme le graphe de Reeb [19, 30, 40], le complexe de Morse-Smale [16, 21, 27], etc...), l'homologie persistante [13, 14] est un outil fondamental qui propose de mesurer l'importance des structures topologiques (composantes connexes, cycles, cavités, etc.) selon leur *durée de vie* (leur plage de valeur dans les données). Cette théorie permet d'introduire une mesure de bruit sur les structures topologiques, appelée *persistance*, dont la stabilité a été démontrée d'un point de vue théorique [10], et qui permet en pratique de distinguer avec efficacité et précision les structures d'intérêt du bruit. L'efficacité pratique de l'homologie persistante a été documentée dans de nombreuses applications, comme en imagerie médicale [2, 6], en biologie cellulaire [22], en mécanique des fluides [9, 18, 24, 36], en physique des matériaux [15, 23, 26], en combustion [4, 5, 20, 25], en chimie moléculaire [3, 17], en astrophysique [32, 33], en traitement de surfaces [37, 39, 41, 42] ou encore en monitoring de simulations numériques haute-performance [31].

2 Problème scientifique

Les données considérées sont typiquement représentées sous la forme d'une fonction scalaire linéaire par morceaux $f : \mathcal{M} \rightarrow \mathbb{R}$, associant une valeur réelle à chaque sommet d'une triangulation $\mathcal{M}$, qui représente un objet géométrique 2D ou 3D. En pratique, f représente dans les applications des niveaux de concentrations [15], des potentiels [17], des intensités [2], des températures [5], etc. Les sous-ensembles de niveau $f^{-1}_{-\infty}(i)$ sont définis comme la pré-image de l'intervalle ouvert $]-\infty, i[$ sur $\mathcal{M}$. Simplement, il s'agit de l'ensemble des points de l'objet au dessous d'une certaine valeur i. Quand i augmente, $f^{-1}_{-\infty}(i)$ change de topologie en un nombre fini de configurations: ses nombres de Betti [43] (nombres de composantes connexes, de cycles indépendants, de cavités, etc...) changent sur des points singuliers, appelés points critiques (sphères de couleur, Fig. 1). Chaque structure topologique de $f^{-1}_{-\infty}(i)$ est donc créée sur un premier point critique à une valeur i, puis détruite sur un second point critique à une valeur $j > i$. Le diagramme de persistance $\mathcal{D}(f)$ [10, 14] (Fig. 1) est une représentation graphique de ce processus, où chaque classe d'homologie persistante (chaque structure topologique) est représentée par une barre verticale pour laquelle la cordonnée en abscisse correspond à la valeur i et les extrémités en ordonnées correspondent à i et j. La *persistance* de la classe est donnée par $|j-i|$. Dans ce diagramme, le bruit topologique apparait donc sous la forme de petites barres, proche de la diagonale ($|j-i| \rightarrow 0$), voir [34].

Dans ce stage, nous souhaitons mettre au point un algorithme de clustering d'ensemble de diagrammes de persistance, afin de regrouper des diagrammes similaires (et donc des jeux de données similaires) au sein d'une même classe. Un ingrédient essentiel des algorithmes de clustering est le calcul du barycentre d'un ensemble. Or, il n'existe pas de manière établie de calculer un barycentre entre plusieurs diagrammes de persistance (de calculer un diagramme de persistance *moyen*).

Parmi les pistes possibles, nous souhaitons étendre des résultats récents des densités de probabilités vers les diagrammes de persistance, notamment avec la notion de barycentre de Wassertein [12].

Dans un diagramme, la distance entre deux points a et b (deux sphères, Fig. 1) peut être donnée par la norme L_p (où 2 est une valeur

• *Julien Tierny is with Sorbonne Université, CNRS, LIP6 UMR 7606, France. E-mails: julien.tierny@lip6.fr.*

usuelle du paramètre p):

$$d_p(a,b) = \sqrt[p]{(a_x - b_x)^p + (a_y - b_y)^p} \quad (1)$$

Ensuite, la distance de Wassertein entre deux diagrammes $\mathscr{D}(f)$ et $\mathscr{D}(g)$, notée d_p^W, est donnée par:

$$d_p^W\left(\mathscr{D}(f),\mathscr{D}(g)\right) = \min_{\phi\in\Phi} \sqrt[p]{\sum_{x\in\mathscr{D}(f)} \left(d_p\left(x,\phi(x)\right)\right)^p} \quad (2)$$

où Φ représente l'ensemble de toutes les mises en correspondance possibles entre les structures du diagramme $\mathscr{D}(f)$ et celle de $\mathscr{D}(g)$. Simplement, cette mesure repose sur une mise en correspondance optimale entre les barres des diagrammes, et mesure ensuite la distance entre diagrammes comme la somme des distances entre barres mises en correspondances par ϕ.

Etant donné un ensemble de diagrammes de persistance $\mathbb{D} = \{\mathscr{D}(f_0),\mathscr{D}(f_1),\ldots,\mathscr{D}(f_n)\}$, le barycentre de Wasserstein $\mathscr{D}(f)^*$ de $\mathbb{D}$ pourrait être introduit comme la moyenne de Fréchet de la distance de Wasserstein:

$$\mathscr{D}(f)^* = \operatorname*{arg\,min}_{\mathscr{D}(f)\in\mathbb{P}} \sum_{\mathscr{D}(f_i)\in\mathbb{D}} d_p^W\left(\mathscr{D}(f),\mathscr{D}(f_i)\right) \quad (3)$$

où $\mathbb{P}$ représente l'ensemble de tous les diagrammes de persistance possibles. Simplement, le barycentre de Wasserstein est défini ici comme un diagramme étant au plus proche, au sens de la distance de Wasserstein, de tous les diagrammes de $\mathbb{D}$. Pour résoudre ce problème d'optimisation difficile, nous proposons d'étendre les travaux sur les densités de probabilités [12] aux diagrammes de persistance, notamment en considérant des techniques de quantification [1].

3 Perspectives

Ce stage est proposé dans l'optique d'une poursuite en thèse de doctorat sur le thème de l'analyse topologique de données haute performance. Ce sujet a été sélectionné comme thème possible pour une thèse CIFRE (co-encadrée par Julien Tierny et Julien Jomier, démarrage prévu fin 2018) dans le cadre d'un partenariat entre l'UPMC et Kitware [47], une société majeure dans le logiciel open-source (CMake [45], CDash [44], VTK [49], ITK [46], ParaView [48]) et la *data science*.

De manière plus générale, ce stage et sa possible poursuite en thèse apporteront un bagage de compétences scientifiques et techniques pointues et recherchées dans le domaine de la *data science* et de l'analyse et de la visualisation interactive de données scientifiques (TDA, TTK [38], ParaView [48]). Il constitue donc une expérience fortement valorisable pour accéder à des fonctions R&D sur ces thèmes, dans le monde académique comme industriel (Kitware, EDF, Total, CEA, etc.).

4 Organisation du stage

Le stage pourra se dérouler selon les étapes suivantes:

1. Etudier la bibliographie existante sur:
   - l'analyse topologique de données [13, 34];
   - les distances et barycentres de Wasserstein pour les densités de probabilité [11, 12];
   - la quantification de diagrammes de persistance [1].
2. Imaginer et mettre en oeuvre un algorithme efficace de calcul de barycentres entre diagrammes de persistance;
3. Valider l'approche d'un point de vue expérimental, dans des applications de *clustering* de données d'ensembles, sur une variété de jeux de données pratiques provenant de divers contextes applicatifs.

Les programmes d'expérimentation seront écrits en C++, sous la forme de modules pour la plate-forme open-source d'analyse topologique de données *"Topology ToolKit"* (TTK) [38] (intégrée à ParaView [48]).

Le stage peut durer de 16 à 24 semaines, selon les disponibilités du stagiaire. Il s'agit d'un stage rémunéré (rémunération académique standard, environ 500 euros par mois).

5 Profil

Nous recherchons un(e) étudiant(e) très motivé(e)! Curiosité, ouverture d'esprit, créativité, et ténacité sont les aptitudes de caractère que nous recherchons. Ce stage s'adresse aux étudiants en dernière année de master en informatique ou mathématiques appliquées (et domaines connexes) ou aux étudiants en dernière année d'école d'ingénieurs. Le stagiaire devra être à l'aise avec la programmation en C++, ou motivé pour le devenir. Un intérêt pour la 3D, la géométrie, la topologie et plus généralement pour les mathématiques et l'informatique est requis.

6 Lieu

Ce stage aura lieu au sein du département Calcul Scientifique du laboratoire d'informatique (LIP6) de Sorbonne Université, en plein coeur de Paris (arrêt Jussieu, lignes 7 et 10). Il sera encadré par Julien Tierny, chercheur au CNRS, expert en analyse topologique de données pour la visualisation et l'analyse de données scientifiques (http://lip6.fr/Julien.Tierny).

7 Candidatures

Nous invitons les candidat(e)s à nous faire parvenir leur lettre de candidature accompagné d'un CV mis à jour à Julien Tierny (julien.tierny@lip6.fr). Nous vous encourageons à nous contacter par email pour toute question ou pour discuter davantage du sujet.

References

[1] H. Adams, S. Chepushtanova, T. Emerson, E. Hanson, M. Kirby, F. Motta, R. Neville, C. Peterson, P. Shipman, and L. Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 2017.

[2] A. Bock, H. Doraiswamy, A. Summers, and C. Silva. Topoangler: Interactive topology-based extraction of fishes. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2017.

[3] R. A. Boto, J. C. Garcia, J. Tierny, and J.-P. Piquemal. Interpretation of the reduced density gradient. *Molecular Physics*, 2016.

[4] P. Bremer, G. Weber, J. Tierny, V. Pascucci, M. Day, and J. Bell. A topological framework for the interactive exploration of large scale turbulent combustion. In *Proc. of IEEE eScience*, 2009.

[5] P. Bremer, G. Weber, J. Tierny, V. Pascucci, M. Day, and J. Bell. Interactive exploration and analysis of large scale simulations using topology-based data segmentation. *IEEE Transactions on Visualization and Computer Graphics*, 2011.

[6] H. Carr, J. Snoeyink, and M. van de Panne. Simplifying flexible isosurfaces using local geometric measures. In *IEEE VIS*, 2004.

[7] F. Chazal, L. Guibas, S. Oudot, and P. Skraba. Persistence-based clustering in Riemannian manifolds. *Journal of the ACM*, 2013.

[8] F. Chazal and J. Tierny. Topological data analysis, online class. `http://lip6.fr/Julien.Tierny/topologicalDataAnalysisClass.html`.

[9] F. Chen, H. Obermaier, H. Hagen, B. Hamann, J. Tierny, and V. Pascucci. Topology analysis of time-dependent multi-fluid data using the reeb graph. *Computer Aided Geometric Design*, 2013.

[10] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. In *Symp. on Comp. Geom.*, 2005.

[11] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proc. of NIPS*, 2013.

[12] M. Cuturi and A. Doucet. Fast computation of wasserstein barycenters. In *Proc. of ICML*, 2014.

[13] H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2009.

[14] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Disc. Compu. Geom.*, 2002.

[15] G. Favelier, C. Gueunet, and J. Tierny. Visualizing ensembles of viscous fingers. In *IEEE SciVis Contest*, 2016.

[16] R. Forman. A user's guide to discrete Morse theory. *Adv. in Math.*, 1998.

[17] D. Guenther, R. Alvarez-Boto, J. Contreras-Garcia, J.-P. Piquemal, and J. Tierny. Characterizing molecular interactions in chemical systems. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2014.

[18] D. Guenther, J. Salmon, and J. Tierny. Mandatory critical points of 2D uncertain scalar fields. *Computer Graphics Forum (Proc. of EuroVis)*, 2014.

[19] C. Gueunet, P. Fortin, J. Jomier, and J. Tierny. Contour forests: Fast multi-threaded augmented contour trees. In *IEEE LDAV*, 2016.

[20] A. Gyulassy, P. Bremer, R. Grout, H. Kolla, J. Chen, and V. Pascucci. Stability of dissipation elements: A case study in combustion. *Computer Graphics Forum (Proc. of EuroVis)*, 2014.

[21] A. Gyulassy, P. T. Bremer, B. Hamann, and V. Pascucci. A practical approach to morse-smale complex computation: Scalability and generality. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2008.

[22] A. Gyulassy, D. Guenther, J. A. Levine, J. Tierny, and V. Pascucci. Conforming morse-smale complexes. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2014.

[23] A. Gyulassy, A. Knoll, K. Lau, B. Wang, P. Bremer, M. Papka, L. A. Curtiss, and V. Pascucci. Interstitial and interlayer ion diffusion geometry extraction in graphitic nanosphere battery materials. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2015.

[24] J. Kasten, J. Reininghaus, I. Hotz, and H. Hege. Two-dimensional time-dependent vortex regions based on the acceleration magnitude. *IEEE Transactions on Visualization and Computer Graphics*, 2011.

[25] D. E. Laney, P. Bremer, A. Mascarenhas, P. Miller, and V. Pascucci. Understanding the structure of the turbulent mixing layer in hydrodynamic instabilities. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2006.

[26] J. Lukasczyk, G. Aldrich, M. Steptoe, G. Favelier, C. Gueunet, J. Tierny, R. Maciejewski, B. Hamann, and H. Leitte. Viscous fingering: A topological visual analytic approach. In *Physical Modeling for Virtual Manufacturing Systems and Processes*, 2017.

[27] J. Milnor. *Morse Theory*. Princeton U. Press, 1963.

[28] MyWhyU. A humorous look at the topology of curved space. `https://www.youtube.com/watch?v=p2ofJPh2yMw&list=PL09E9E697F585A58C`.

[29] V. Pascucci, X. Tricoche, H. Hagen, and J. Tierny. *Topological Methods in Data Analysis and Visualization: Theory, Algorithms and Applications*. Springer, 2010.

[30] G. Reeb. Sur les points singuliers d'une forme de Pfaff complètement intégrable ou d'une fonction numérique. *Acad. des Sci.*, 1946.

[31] E. Santos, J. Tierny, A. Khan, B. Grimm, L. Lins, J. Freire, V. Pascucci, C. Silva, S. Klasky, R. Barreto, and N. Podhorszki. Enabling advanced visualization tools in a web-based simulation monitoring system. In *Proc. of IEEE eScience*, 2009.

[32] N. Shivashankar, P. Pranav, V. Natarajan, R. van de Weygaert, E. P. Bos, and S. Rieder. Felix: A topology based framework for visual exploration of cosmic filaments. *IEEE Transactions on Visualization and Computer Graphics*, 2016. `http://vgl.serc.iisc.ernet.in/felix/index.html`.

[33] T. Sousbie. The persistent cosmic web and its filamentary structure: Theory and implementations. *Royal Astronomical Society*, 2011. `http://www2.iap.fr/users/sousbie/web/html/indexd41d.html`.

[34] J. Tierny. Introduction to topological data analysis. `https://hal.archives-ouvertes.fr/cel-01581941/file/manuscript.pdf`.

[35] J. Tierny. *Contributions to Topological Data Analysis for Scientific Visualization*. Habilitation (HDR), Sorbonne University UPMC, 2016.

[36] J. Tierny and H. Carr. Jacobi fiber surfaces for bivariate Reeb space computation. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2016.

[37] J. Tierny, J. Daniels, L. G. Nonato, V. Pascucci, and C. Silva. Interactive quadrangulation with Reeb atlases and connectivity textures. *IEEE Transactions on Visualization and Computer Graphics*, 2012.

[38] J. Tierny, G. Favelier, J. A. Levine, C. Gueunet, and M. Michaux. The Topology ToolKit. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2017. `https://topology-tool-kit.github.io/`.

[39] J. Tierny, D. Guenther, and V. Pascucci. Optimal general simplification of scalar fields on surfaces. In *Topological and Statistical Methods for Complex Data*. Springer, 2014.

[40] J. Tierny, A. Gyulassy, E. Simon, and V. Pascucci. Loop surgery for volumetric meshes: Reeb graphs reduced to contour trees. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2009.

[41] J. Tierny and V. Pascucci. Generalized topological simplification of scalar fields on surfaces. *IEEE Transactions on Visualization and Computer Graphics (Proc. of IEEE VIS)*, 2012.

[42] A. Vintescu, F. Dupont, G. Lavoué, P. Memari, and J. Tierny. Conformal factor persistence for fast hierarchical cone extraction. In *Eurographics (short papers)*, 2017.

[43] Wikipedia. Betti numbers. `https://en.wikipedia.org/wiki/Betti_number`.

[44] Wikipedia. CDash. `https://www.cdash.org/`.

[45] Wikipedia. CMake. `https://en.wikipedia.org/wiki/CMake`.

[46] Wikipedia. ITK. `https://en.wikipedia.org/wiki/Insight_Segmentation_and_Registration_Toolkit`.

[47] Wikipedia. Kitware Inc. `https://en.wikipedia.org/wiki/Kitware`.

[48] Wikipedia. ParaView. `https://en.wikipedia.org/wiki/ParaView`.

[49] Wikipedia. VTK. `https://en.wikipedia.org/wiki/VTK`.