Research Statement: Building Bridges between Statistical Learning and Topological Data Analysis

Mathieu Carrière

The classification and analysis of available data sets has recently become a problem of primary interest to data scientists. During my PhD, I worked on applying statistical and Machine Learning (ML) methods to representations of data coming from methods in Topological Data Analysis (TDA), such as *persistence diagrams* (PDs), *Reeb spaces* and *Mappers* [1]. I defined stable feature maps that send PDs to Hilbert spaces in order to apply kernel methods to PDs. I also analyzed the topological structure of the Mapper depending on its parameters, and derived an appropriate metric to compare Mappers and Reeb spaces. Building on these results, I developed a statistical framework for Mapper, shedding a new light on it that allows us to perform parameter selection, compute confidence regions and prove optimal rates of convergence.

Persistence Diagrams.

Homology theory lies at the core of the PD definition. The shape of a space X can be measured with the so-called *homology groups* [Mun93]. For any dimension k, the corresponding kth homology group of X, denoted $H_k(X)$, is the group whose basis is given by the k-dimensional holes in X (connected components when $k = 0$, loops when $k = 1$, cavities when $k = 2$, etc). PDs can then be computed from a *filtration* of a topological space X, i.e. a sequence of nested subspaces of X: $\emptyset = X_1 \subseteq X_2 \subseteq X_3 \subseteq \ ... \ \subseteq X_n = X$. This definition can be extended to filtrations indexed over the real line [CdSGO16]. A common example of filtration is the sequence of sublevel sets of a continuous function $\{f^{-1}(-\infty, \alpha)\}_{\alpha \in \mathbb{R}}$. By computing homology groups, a filtration leads to a sequence of groups connected by morphisms, also called a *persistence module* [CdSGO16]: $H_*(X_1) \to H_*(X_2) \to H_*(X_3) \to \ ... \ \to H_*(X_n)$. The idea of persistent homology is to look at generators that persist in this sequence, i.e. to track the changes in homology as one goes through the sequence. If a generator appears ("is born") at time $b \subseteq \{1, ..., n\}$ and disappears ("dies") at time $d \subseteq \{1, ..., n\}$, the corresponding PD includes a point with coordinates (b, d). Note that points in a PD always lie above the diagonal since $d \geq b$. See Figure 1 for an illustration.

[1] In this document, we call *Mapper* the mathematical object, not the algorithm used to compute it.

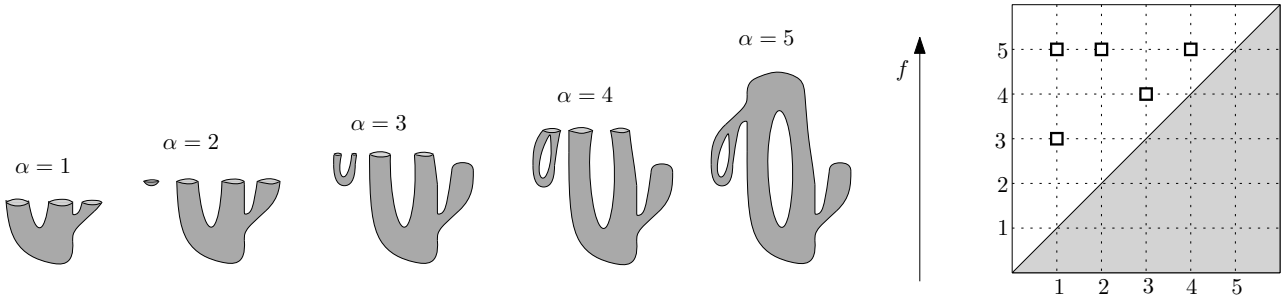

Figure 1: Example of a PD induced by the height function f.

The common metric for PDs, the so-called *bottleneck distance* d_{B}, can be seen as the l_∞-cost of an optimal transportation plan between PDs, in which points are allowed to be snapped onto the diagonal. Considering the l_p-cost, $p \in \mathbb{N}$, leads to the so-called *Wasserstein distance* W_∞^p. A key property justifying the use of PDs is their *stability* [CSEH07]: $d_{\mathrm{B}}(D(f), D(g)) \leq \|f - g\|_\infty$ for any continuous space X and "tame" functions $f, g : X \to \mathbb{R}$.

Kernels Methods. The use of kernel methods allows a user to apply ML algorithms when the objects live in a space X that is not the traditional Euclidean space $\mathbb{R}^d$. This is useful for algorithms that require Hilbert space structure, like SVM or PCA. The idea is to simply send these objects into a Hilbert space H through a *feature map*, where computations are well-defined and easier. If a user does not know H in advance, a useful theorem of Moore and Aronszajn [Aro50] states that for any positive semi-definite function $K : X \times X \to \mathbb{R}$ (a "kernel") there always exists an essentially unique Hilbert space H_K and a corresponding feature map $\Psi_K : X \to H_K$ such that $K(x_1, x_2) = \langle \Psi_K(x_1), \Psi_K(x_2) \rangle_{H_K}$. If (X, d) is a metric space, one of the most common kernel is defined by $K_\sigma(x_1, x_2) = \mathrm{e}^{-\frac{d(x_1,x_2)}{2\sigma^2}}$, and is called a *Gaussian* kernel.

Feature Map for PDs. In [COO15b], a collaboration with Maks Ovsjanikov and Steve Oudot, we provided a feature map Φ from PDs to Euclidean space $\mathbb{R}^d$. This allowed us to directly apply kernels on $\mathbb{R}^d$. Indeed, for any kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $K' : (D_1, D_2) \mapsto K(\Phi(D_1), \Phi(D_2))$ is a valid kernel for PDs. It is worth noting that K' does not use the ambient metric d_{B} for PDs, avoiding possibly costly matching computations. In our case, Φ is obtained by computing all pairwise distances in the PD and by truncating their distribution at a fixed threshold d. We proved in [COO15a] that Φ is stable, i.e. $\|\Phi(D_1) - \Phi(D_2)\| \leq C d_{\mathrm{B}}(D_1, D_2)$, for some $C > 0$ that depends on the norm used. In particular, $C = 2$ for $\|\cdot\|_\infty$ and $C = (N(N-1)/2)^{1/2}$, where $N = \max\{|D_1|, |D_2|\}$, for $\|\cdot\|_2$. We applied this feature map to *local* PDs computed from points on 3D shapes by considering growing geodesic balls centered on the points. We also proved a stability result for such PDs.

Gaussian kernel for PDs. In [CCO17], which I wrote in collaboration with Marco Cuturi and Steve Oudot, we studied the case of Gaussian kernels for PDs. A classical result of Berg

et al. [BCR84] states that K_σ is a valid kernel for all $\sigma > 0$ (i.e. positive semi-definite) if and only if d is *conditionally negative semi-definite* (c.n.d.), i.e. $\sum_{i,j} a_i a_j d(x_i, x_j) \leq 0$, $\forall x_1, ..., x_n$ and $a_1, ..., a_n$ such that $\sum_i a_i = 0$. Unfortunately, the bottleneck distance d_{B} for PDs is not c.n.d., and there is evidence that Wasserstein distances are not either. Thus, the use of d_{B} for kernel methods is not permitted. However, there exists a modification of W_∞^1 called the *Sliced Wasserstein distance* SW that is c.n.d. Intuitively, this distance can be computed by integrating the first Wasserstein distance W_∞^1 between the projections of the PDs onto a line over all lines that include the origin. We proved that this distance is c.n.d., thus leading to a valid Gaussian kernel. Additionally, it preserves the metric i.e. there exist constants $C_1, C_2 > 0$ such that $C_1 W_\infty^1 \leq \mathrm{SW} \leq C_2 W_\infty^1$, in contrast with all other kernels for PDs that only focus on the upper bound. This equivalence helps improving the discriminative power of the kernel, as shown by the large improvements obtained on accuracies in several benchmarks applications.

Reeb spaces and Mappers.

One way of characterizing the structure of a continuous function $f : X \to \mathbb{R}^d$ is to look at the evolution of the topology of its *level sets* $f^{-1}(\{\alpha\})$, for α ranging over $\mathbb{R}^d$. This information is summarized in a mathematical object called the *Reeb space* of a pair (X, f), denoted by $\mathrm{R}_f(X)$ and defined as the quotient space obtained by identifying the points of X that lie in the same connected component of the same level set of f [Ree46]. However, in practice it is costly to compute it exactly. The *Mapper* was introduced by Singh, Mémoli and Carlsson [SMC07] as a computable discretization of Reeb spaces. Its construction depends on the choice of a cover $\mathcal{I}$ of the image of f by open sets. Pulling back $\mathcal{I}$ through f gives an open cover of the domain X. This cover may have some elements that are disconnected, so it is refined into a connected cover by splitting each element into its various connected components. These connected components are well-defined in the case of continuous spaces. In the case of point clouds, they can be computed using clustering. Then, the Mapper $\mathrm{M}_{f,\mathcal{I}}(X)$ is defined as the nerve of the connected cover, having one vertex per element, one edge per pair of intersecting elements, and more generally, one k-simplex per every non-empty $(k+1)$-fold intersection. See Figure 2 for an example.

Structural Analysis of Mapper. In [CO16], a collaboration with Steve Oudot, I studied the influence of the cover $\mathcal{I}$ on the topology of $\mathrm{M}_{f,\mathcal{I}}(X)$, when f is scalar. We proved that the topological structure of the Mapper is a simplification of that of the Reeb space $\mathrm{R}_f(X)$ (also called Reeb graph when f is scalar). We showed that the missing topological features can be deduced from the cover by deriving a subset of the plane, called *staircase*, which, when applied to PDs, show the points that disappear when going from the Reeb graph to the Mapper. These staircases allowed us to define a slight generalization $d_{\mathcal{I}}$ of the bottleneck distance that take covers into account, thus providing a natural metric to compare Mappers computed on topological spaces. Finally, we adapted this framework for point clouds in order to be able to do statistics.

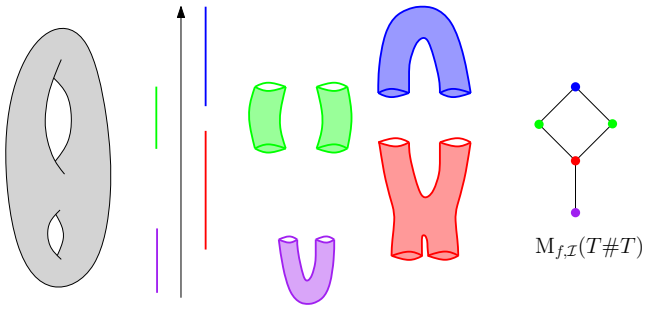

Figure 2: Example of the Mapper computed on a double torus with the height function and a cover given by four different open intervals.

Metric Equivalence for Reeb graphs. In [CO17], a collaboration with Steve Oudot, we justified the use of the bottleneck distance d_{B} as a natural metric to compare Reeb graphs (and $d_{\mathcal{I}}$ as a natural choice for Mappers) when functions are scalars, by showing that d_{B} is *locally equivalent* to the other existing distances d_{FD} and d_{I}. In particular, this allowed us to show that their *induced metrics* [BBI01] $\hat{d}_{\mathrm{B}}$, $\hat{d}_{\mathrm{FD}}$ and $\hat{d}_{\mathrm{I}}$ are all *globally equivalent* in the sense that there exist constants $C_1, C_2 > 0$ such that $C_1\hat{d} \leq \hat{d}_{\mathrm{B}} \leq C_2\hat{d}$, where d is either d_{FD} or d_{I}, suggesting that $\hat{d}_{\mathrm{B}}$ is a canonical metric for Reeb graphs. Finally, we provided few remarks on the space of Reeb graphs equipped with $\hat{d}_{\mathrm{B}}$, such as non-completeness, and emphasized its connection with the problem of finding interpolations and barycenters for Reeb graphs.

Statistical Analysis of Mapper. In [CMO17], which is a collaboration with Bertrand Michel and Steve Oudot currently submitted to Journal of Machine Learning Research, we built on the previous works to study Mapper from a statistical point of view. Indeed, using d_{B} as a metric between $\mathrm{M}_{f,\mathcal{I}}(X)$ and $\mathrm{R}_f(X)$, we proved that the Mapper computed on a point cloud with specific parameter $\hat{\mathcal{I}}$ is a minimax optimal estimator of the Reeb graph. We also showed empirically that $\hat{\mathcal{I}}$ is a good candidate for parameter tuning. Moreover, the discrete results obtained in the last section of [CO16] were used to define approximation inequalities that allowed us to compute confidence regions, i.e. quantities such as $\mathrm{Proba}(d_{\mathrm{B}}(\mathrm{M}_{f,\hat{\mathcal{I}}}(X), \mathrm{R}_f(X)) > \alpha)$ for some $\alpha \in \mathbb{R}$. We also provided empirical evidence on the efficiency of *bootstrap* methods to compute such regions.

Future Directions

Extension to general spaces. The structural analysis of Mapper is an open problem when the filter function is multivariate, and so is the local equivalence of metrics when one studies general topological spaces instead of Reeb graphs. I plan to work on these extensions since they would strongly strenghten the results.

Bootstrap Validity of the Mapper. When doing the statistical analysis of the Mapper computed on real data in the work mentioned above, the best confidence bands were obtained

using bootstrap. However, this efficiency is only empirical and proving that bootstrap is valid requires more work. It has been done already for PDs [CFL$^+$14] through a careful analysis. Since checking bootstrap validity for Mappers would be very useful for practical applications, I would like to carry out such an analysis. I also plan to study generalizations of kernels on graphs [VSKB10] to kernels on simplicial complexes, using i.e. random walks on simplicial complexes [MS13], in order to use Mappers in ML problems.

Metric Geometry Properties of the Space of Reeb Graphs. In [CO17], we emphasized the advantages of studying the space of Reeb graphs, since it would lead to a deeper understanding of many practical questions concerning these objects, such as the definition of geodesics, interpolates and barycenters. In [CO17], we pointed out the technical difficulties when dealing with such questions, especially when compared to their analogues in the space of PDs [TMMH14]. This justifies the need for a careful study that I would like to work on.

Survey of Kernels for PDs. Many vectorization methods and kernels for PDs have been defined and studied over the past few years [RHBK15, dFF15, ACE$^+$15, RT16, Ver16, KFH16]. They all differ in their definitions, properties and practical results. I would like to work on a fair and exhaustive comparison, that is still missing in the literature even though it would be very useful for practitioners. Moreover, I think that studying kernel mean embedding applications [MFSS16] for distributions of PDs is of great interest, especially with the recent use of combinations of *local* PDs [COO15a] to describe shapes.

References

[ACE$^+$15] Henry Adams, Sofya Chepushtanova, Tegan Emerson, Eric Hanson, Michael Kirby, Francis Motta, Rachel Neville, Chris Peterson, Patrick Shipman, and Lori Ziegelmeier. Persistent Images: A Stable Vector Representation of Persistent Homology. *CoRR*, abs/1507.06217, 2015.

[Aro50] Nachman Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[BBI01] Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A Course in Metric Geometry*. American Mathematical Society, 2001.

[BCR84] Christian Berg, Jens Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Springer, 1984.

[CCO17] Mathieu Carrière, Marco Cuturi, and Steve Oudot. Sliced Wasserstein Kernel for Persistence Diagrams. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[CdSGO16] Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. *The Structure and Stability of Persistence Modules*. Springer, 2016.

[CFL+14] Frédéric Chazal, Brittany Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Robust topological inference: distance to a measure and kernel distance. *CoRR*, abs/1412.7197, 2014. Accepted for publication in Journal of Machine Learning Research.

[CMO17] Mathieu Carrière, Bertrand Michel, and Steve Oudot. Statistical Analysis and Parameter Selection for Mapper. *CoRR*, abs/1706.00204, 2017.

[CO16] Mathieu Carrière and Steve Oudot. Structure and Stability of the 1-Dimensional Mapper. In *Proceedings of the 32nd Symposium on Computational Geometry*, volume 51, pages 25:1–25:16, 2016.

[CO17] Mathieu Carrière and Steve Oudot. Local Equivalence and Induced Metrics for Reeb Graphs. In *Proceedings of the 33rd Symposium on Computational Geometry*, 2017.

[COO15a] Mathieu Carrière, Steve Oudot, and Maks Ovsjanikov. Local Signatures using Persistence Diagrams. *HAL preprint*, 2015.

[COO15b] Mathieu Carrière, Steve Oudot, and Maks Ovsjanikov. Stable Topological Signatures for Points on 3D Shapes. *Computer Graphics Forum*, 34, 2015.

[CSEH07] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of Persistence Diagrams. *Discrete and Computational Geometry*, 37(1):103–120, 2007.

[dFF15] Barbara di Fabio and Massimo Ferri. Comparing persistence diagrams through complex vectors. *CoRR*, abs/1505.01335, 2015.

[KFH16] Genki Kusano, Kenji Fukumizu, and Yasuaki Hiraoka. Persistence Weighted Gaussian Kernel for Topological Data Analysis. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2004–2013, 2016.

[MFSS16] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel Mean Embedding of Distributions: A Review and Beyond. *CoRR*, abs/1605.09522, 2016.

[MS13] Sayan Mukherjee and John Steenbergen. Random Walks on Simplicial Complexes and Harmonics. *Random Structures and Algorithms*, 48(2), 2013.

[Mun93] James Munkres. *Elements of Algebraic Topology*. Addison-Wesley, 1993.

[Ree46] Georges Reeb. Sur les points singuliers d'une forme de Pfaff complètement intégrable ou d'une fonction numérique. *Compte Rendu de l'Académie des Science de Paris*, 222:847–849, 1946.

[RHBK15] Jan Reininghaus, Stefan Huber, Ulrich Bauer, and Roland Kwitt. A Stable Multi-Scale Kernel for Topological Machine Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[RT16] Vanessa Robins and Katharine Turner. Principal Component Analysis of Persistent Homology Rank Functions with case studies of Spatial Point Patterns, Sphere Packing and Colloids. *Physica D: Nonlinear Phenomena*, 334:1–186, 2016.

[SMC07] Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In *Symposium on Point Based Graphics*, pages 91–100, 2007.

[TMMH14] Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet Means for Distributions of Persistence Diagrams. *Discrete and Computational Geometry*, 52(1):44–70, 2014.

[Ver16] Sara Kalisnik Verovsek. Tropical Coordinates on the Space of Persistence Barcodes. *CoRR*, abs/1604.00113, 2016.

[VSKB10] S. V. N. Vishwanathan, Nicol Schraudolph, Risi Kondor, and Karsten Borgwardt. Graph Kernels. *Journal of Machine Learning Research*, 11:1201–1242, 2010.