An Optimal Partial Transport perspective on Topological Descriptors

Théo Lacombe

Topological Data Analysis (TDA) provides a machinery to extract and summarize topological information from complex structured objects; think of graphs, points sampled on a manifold, or time series for instance. During my PhD, I worked on developing new theoretical and numerical tools to study the most common topological descriptor: the *persistence diagram* (PD), targeting statistical and learning applications. To do so, I studied and clarified the apparent links between the widely developed Optimal Transport (OT) theory and TDA. Primarily, I showed how PD metrics can be formulated as optimal *partial* transport problems, shedding a new light on the understanding of the space of PDs and allowing to adapt many theoretical results from OT to PDs. Building on recent advances in computational OT, I developed efficient algorithms to deal with PDs, in particular to estimate barycenters of large samples of PDs.

Bridging OT and TDA

Background

Optimal Transport. Consider two probability measures μ, ν supported on some Polish space Ω, along with a cost function $c : \Omega \times \Omega \to \mathbb{R}_+$. Informally, one can interpret μ as an initial distribution of mass, ν as a target distribution (which obviously must have the same mass as μ), while $c(x, y)$ is the cost of transporting a unit of mass from x to y. In its standard formulation, OT seeks for the best way of transporting μ onto ν by looking for a measure π supported on $\Omega \times \Omega$ with marginals μ, ν that would minimize $\iint c(x, y) \mathrm{d}\pi(x, y)$. When c is of the form $d(x, y)^p$, where d is a metric on Ω and $1 \leq p < \infty$, the p-th root of the optimal transport cost that can be achieved is called the p-*Wasserstein distance* between μ and ν, written $W_p(\mu, \nu)$, and the corresponding metric space $(\mathcal{W}^p(\Omega), W_p)$ is referred to as the Wasserstein space on Ω. This space has been extensively studied, from both theoretical [30, 26] and computational [24] perspectives.

Persistent Homology and Persistence Diagrams. Persistence Diagrams aim to summarize the topology of an object in a multi-scale fashion. Given a topological space X and a real-valued function $f : X \to \mathbb{R}$, the t-sublevel set of (X, f) is defined as $X_t :=$

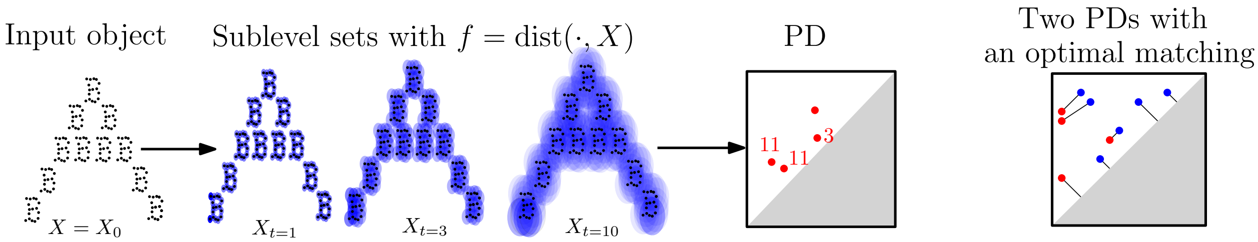

Figure 1: *(left)* A sketch of the TDA pipeline, for X a point cloud and f the distance to the compact. Observe that X_t is the union of balls centered at $x \in X$ of radius t. Points in the persistence diagram records appearance and disappearance of loops (1-dimensional topological features) in this union of balls. *(right)* Two PDs μ and ν and an optimal matching between them. The distance $d_p(\mu, \nu)$ is defined as the p-th root of the sum of the lengths to the p of all edges appearing in the matching.

$f^{-1}((-\infty, t]) = \{x \in X,\ f(x) \leq t\}$. Making t increase from $-\infty$ to $+\infty$ gives an increasing sequence of topological spaces, called the *filtration* of X by f. It starts with the empty set and ends with the whole space X. Informally, persistent homology [11, 23] will track in $(X_t)_t$ the scales of appearance and disappearance of topological features (connected component, loops, cavities, etc.). For instance, a loop (a "one-dimensional topological feature") might appear in X_{t_b} at a given scale t_b, called its *birth time*, and eventually disappear (get "filled") at scale $t_d > t_b$, its *death time*. One says that the loop *persists* on the interval $[t_b, t_d]$. The family of all such intervals is called the *persistence diagram* of (X, f) and can be represented as a point cloud supported on the upper half-plane $\Omega := \{(t_1, t_2),\ t_2 > t_1\} \subset \mathbb{R}^2$, where points can appear with some (finite) multiplicity, see Figure 1. Equivalently, one can represent a persistence diagram as a *point measure*, that is a Radon measure[1] of the form $\sum_{x \in P} n_x \delta_x$, where P is a locally finite subset of Ω, $n_x \in \mathbb{N}$, and δ_x denotes the Dirac mass at x.

Statistics with PDs. Now, consider a set of observed objects $X_1 \ldots X_n$ (or, more generally, a distribution $\boldsymbol{X}$ of such objects), and their respective PDs $\mu_1 \ldots \mu_n$ (respectively, a distribution of diagrams $\boldsymbol{\mu}$). One could be interested in using the diagrams μ_is to get new statistical descriptors on the input sample with a topological flavor. For instance, one may want to describe the "average topology" of these observations, which turns out to compute a *barycenter* in the space of PDs, and leads to the more general question: *how to perform statistics with persistence diagrams?* The space of PDs is not a Hilbert space but only a metric space. The distance d_p between two diagrams μ, ν (where $1 \leq p \leq \infty$) is defined as the minimum transport cost (with $c(x, y) = \|x - y\|^p$) to match points in μ onto either points in ν (in a one-to-one way) or onto the diagonal $\partial\Omega = \{t_1 = t_2\}$ (and symmetrically, points in ν that are not matched with a point in μ must also be matched to the diagonal), see Figure 1. Statistical properties of the metric space $(\mathcal{D}, d_p)$ have been studied in the seminal papers [22, 29].

[1] A Radon measure is a locally finite Borel measure.

Contributions

An OT formulation of PD metrics. Obviously, d_p and W_p metrics share the key idea of matching distributions of masses. This similarity is known to the TDA-community for long, to such a point that the d_p-metrics are sometimes referred to as "*Wasserstein distances between PDs*". However, the peculiar role played by the diagonal $\partial\Omega$ (which allows in particular for difference of masses in diagrams) refrained from going further into this similarity. In [4], Carriere et al. however showed how the *Sliced-Wasserstein kernel*, a standard tool in OT, can be successfully adapted to handle PDs, suggesting further connections between those fields.

In [10], a collaboration[2] with Vincent Divol, building on a work of Figalli and Gigli [14], we actually proved that the metrics d_p could be expressed as particular cases of Optimal *Partial* Transport problem. This reformulation has multiple strengths: first, it makes sense for a larger class of measures, that we call *persistence measures*, than just persistence diagrams (which are discrete by nature). Continuous counterpart of persistence diagrams naturally arise in random settings [5], and the combinatorial definition of the d_p metrics is not suited to handle measures with a non-discrete supports and non-uniform mass distributions. In contrast, our formulation allows us to prove new results concerning persistence diagrams. Notably,

- Characterization of the maps $f : \Omega \to \mathcal{B}$ for some Banach space $(\mathcal{B}, \|\cdot\|)$ such that the *linear vectorization*[3] $\mu \mapsto \mu(f)$ is continuous from $(\mathcal{D}, d_p)$ to $\mathcal{B}, \|\cdot\|)$.
- Existence of Fréchet means (that is, barycenter) for any probability distribution supported on $(\mathcal{D}, d_p)$, extending the results of [29].
- Topological stability of random process, that is of the map $\xi \mapsto \mathbb{E}_{\xi^{\otimes n}}[\mathrm{Dgm}(X)]$, where ξ is a probability measure, X is a n-sample of law ξ, $\mathrm{Dgm}(X)$ is its Čech diagram, and $\mathbb{E}[\mu]$ for some persistence diagram μ is defined as $\mathbb{E}[\mu](K) = \mathbb{E}[\mu(K)]$ for $K \subset \Omega$.

Efficient algorithms for TDA. Aside theoretical properties, this reformulation legitimates the adaptation of Computational OT tools [24] to deal with PDs. In [20], a collaboration[4] with Marco Cuturi and Steve Oudot, we tackled the problem of estimating the Fréchet mean of a finite sample of persistence diagrams. Given a set of observed PDs $b_1 \dots b_n$, a Fréchet mean of $(b_i)_i$ is a minimizer of

$$a \mapsto \sum_{i=1}^{n} d_2(a, b_i). \tag{1}$$

Although an algorithm to estimate a minimizer of (1) is proposed in [29], it is not convex (and falls into arbitrary bad local minima) and does not scale on large samples. To improve

[2] Submitted at the Journal of Foundation of Computational Mathematics

[3] A formalism that encompasses most of vectorization methods for PDs.

[4] published at NeurIPS 2018

on this, we leverage *regularized* OT [7], where we approximate $d_2(a, b_i)$ by a quantity $d_2^\gamma(a, b_i)$ (for some regularization parameter $\gamma > 0$). The (regularized) map

$$a \mapsto \sum_i d_2^\gamma(a, b_i) \tag{2}$$

is differentiable and we showed how it and its gradient can be expressed thanks to the ("dual") variable (u, v) that is a fixed point of the *Sinkhorn map* $S : (u, v) \mapsto S(u, v)$ (see [7, 20] for details). Such a fixed point is in practice found by "sufficiently" iterating $(u_{t+1}, v_{t+1}) \leftarrow S(u_t, v_t)$ for any initial (u_0, v_0). We show that, even taking the diagonal into account, this operation remains parallelizable and GPU-friendly, thus able to provide estimate in large-scale settings (say, thousands of PDs with thousands of point each). Adopting an Eulerian approach, the (regularized) map (2) is convex, providing a simple gradient descent algorithm to estimate barycenters in the PD-space. In order to control the error made by the Sinkhorn algorithm (i.e. iterating S), we also provide a routine to get *on-the-fly* upper and lower bounds on the error made, that is we provide algorithms to compute bounds m_t^γ and M_t^γ such that after t iterations of the Sinkhorn map with smoothing parameter γ, one has

$$\forall i, m_t^\gamma \leq d_2(a, b_i) \leq M_t^\gamma, \tag{3}$$

while $|M_t^\gamma - m_t^\gamma| \to 0$ as $t \to \infty, \gamma \to 0$.

In a similar vein, other tools of (regularized) OT can be transposed to handle PDs: distance estimation, quantization, differentiability, etc. [8, 17]. Those are the purpose of ongoing work. Some of these algorithms have been or will be integrated to the `Gudhi` library [28].

A theoretical framework to regularize PD metrics. While the primary motivation to introduce regularized OT belongs in its appealing numerical strengths, recent works showed that it is also theoretically founded. In particular, the γ-*Sinkhorn divergence* $S_p^\gamma(\mu, \nu)$ between two probability measures μ and ν interpolates between $W_p(\mu, \nu)$ and the so-called energy distance $\mathrm{ED}(\mu, \nu)$ as γ goes to 0 and $+\infty$ respectively [17, 25]. Furthermore, for any $\gamma > 0$, S_p^γ induces the same topology as W_p does [12]. Extending these results to the PD space is however challenging, in particular as diagrams can have different (even infinite) masses. This problem is the purpose of current work that is expected to be submitted in following weeks as of the day I am writing these lines.

A neural network layer for PDs. In [3], a collaboration[5] with Mathieu Carrière, Frederic Chazal, Yuichi Ike, Martin Royer and Yuhei Umeda, we proposed a unified framework to incorporate PDs in learning pipelines. Our formulation encompasses most of common vectorizations of persistence diagrams [1, 2, 6, 19] in a learnable way, with theoretical guarantees if needed. As we showcase our approach on a graph classification task, we also introduce in this work a new class of topological features on graph: the extended persistence of the Heat Kernel Signature, for which we prove stability properties.

[5] Submitted at AISTATS 2020.

Further directions

Understanding continuous counterpart of PDs. PDs are intrinsically defined as discrete measures with integer mass on each point of their support. However, more general measures (e.g. with a continuous support) can arise as one consider PDs coming from random processes. The bridge we built in [10] gives us the tools to study such measures and to address new problems in TDA. For instance, one can approximate a PD coming from some object X by a continuous measure $\boldsymbol{\mu}$ in a now quantifiable way (e.g. by convolution with a Gaussian): can we "invert" this approximation, that is provide a random object $\boldsymbol{X}$ whose random PD would admit $\boldsymbol{\mu}$ as density? Is $\boldsymbol{X}$ close to the deterministic measure δ_X in the Wasserstein space? Similarly, the geodesic between two continuous persistence measures is *uniquely* defined (which is not the case with discrete measures in general). Does it make sens to "interpolate back" between the generating random processes, similarly to [15]? Intuitively, allowing for more general measures is likely to make the TDA-pipeline more flexible and more suited for probabilistic and statistical analysis.

Statistics in partial-OT spaces. Most of the results presented in [10] are not specific to the PD space and remain valid in the general Optimal *Partial* Transport setting as introduced in [14].[6] Studying the geometry of these partial-OT spaces is of interest as it can lead to new statistical results. For instance, [18] proved that given a probability distribution $\boldsymbol{\mu}$ supported on the Wasserstein space $(\mathcal{W}_2(\Omega), W_2)$ with barycenter $\boldsymbol{b}$ and a n-sample $\mu_1, \dots, \mu_n \sim \boldsymbol{\mu}$ with barycenter b_n, one has $\mathbb{E}[W_2(b_n, \boldsymbol{b})] \leq C/n$, for some constant C and under some regularity assumptions on the transport plans between $\boldsymbol{b}$ and $\mu \in \mathrm{spt}(\boldsymbol{\mu})$. It is very likely that a similar result holds in partial-OT spaces (in particular in the PD space), providing we find the proper way to adapt those assumptions. Similar questions could be addressed, such as convergence rates of empirical measures [31], etc.

Geometry of regularized OT spaces. The Sinkhorn S_p^γ divergence mentioned above is not a metric on the Wasserstein space $\mathcal{W}_p(\Omega)$ as it does not satisfy the triangle inequality in general (it however satisfies the other metric axioms and metricizes the same topology as long as Ω is compact [12]). We could however build a metric from it by simply writing $W_p^\gamma(\mu, \nu) := \min(S_p^\gamma(\mu, \nu), \inf_\lambda \{S_p^\gamma(\mu, \lambda) + S_p^\gamma(\lambda, \nu)\})$ (e.g. is it still non-negatively curved?). It could be interesting to understand how the geometry of $(\mathcal{W}_p(\Omega), W_p^\gamma)$ changes w.r.t. the regularization parameter γ. This would lead to a better understanding of statistical and learning properties of these metrics: behavior of smoothed Wasserstein barycenters [27, 9], sample complexities [16], interpolation and geodesic shooting (intrinsically linked with the structure of tangent cones [21]), etc. Obviously, one can extend this problematic to the regularization of partial-OT spaces mentioned above.

[6] Note that this formulation differs from the one introduced by one of the author in [13], as it allows to transport mass onto the boundary of the space, providing we pay the transportation cost.

References

[1] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: a stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8):1–35, 2017.

[2] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16(1):77–102, 2015.

[3] Mathieu Carrière, Frédéric Chazal, Yuichi Ike, Théo Lacombe, Martin Royer, and Yuhei Umeda. A general neural network architecture for persistence diagrams and graph classification. *arXiv preprint arXiv:1904.09378*, 2019.

[4] Mathieu Carrière, Marco Cuturi, and Steve Oudot. Sliced wasserstein kernel for persistence diagrams. In *34th International Conference on Machine Learning*, 2017.

[5] Frédéric Chazal and Vincent Divol. The Density of Expected Persistence Diagrams and its Kernel Based Estimation. In Bettina Speckmann and Csaba D. Tóth, editors, *34th International Symposium on Computational Geometry (SoCG 2018)*, volume 99 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 26:1–26:15, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[6] Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. Stochastic convergence of persistence landscapes and silhouettes. In *Proceedings of the thirtieth annual symposium on Computational geometry*, page 474. ACM, 2014.

[7] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[8] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

[9] Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

[10] Vincent Divol and Théo Lacombe. Understanding the topology and the geometry of the persistence diagram space via optimal partial transport. *arXiv preprint arXiv:1901.03048*, 2019.

[11] Herbert Edelsbrunner and John Harer. *Computational topology: an introduction.* American Mathematical Soc., 2010.

[12] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-Ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. *arXiv preprint arXiv:1810.08278*, 2018.

[13] Alessio Figalli. The optimal partial transport problem. *Archive for rational mechanics and analysis*, 195(2):533–560, 2010.

[14] Alessio Figalli and Nicola Gigli. A new transportation distance between non-negative measures, with applications to gradients flows with dirichlet boundary conditions. *Journal de mathématiques pures et appliquées*, 94(2):107–130, 2010.

[15] Marcio Gameiro, Yasuaki Hiraoka, and Ippei Obayashi. Continuation of point clouds via persistence diagrams. *Physica D: Nonlinear Phenomena*, 334:118–132, 2016.

[16] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. *arXiv preprint arXiv:1810.02733*, 2018.

[17] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. *arXiv preprint arXiv:1706.00292*, 2017.

[18] Thibaut Le Gouic, Quentin Paris, Philippe Rigollet, and Austin J Stromme. Fast convergence of empirical barycenters in alexandrov spaces and the wasserstein space. *arXiv preprint arXiv:1908.00828*, 2019.

[19] Christoph Hofer, Roland Kwitt, Marc Niethammer, and Andreas Uhl. Deep learning with topological signatures. In *Advances in Neural Information Processing Systems*, pages 1634–1644, 2017.

[20] Théo Lacombe, Marco Cuturi, and Steve Oudot. Large scale computation of means and clusters for persistence diagrams using optimal transport. In *Advances in Neural Information Processing Systems*, 2018.

[21] John Lott. On tangent cones in wasserstein space. *Proceedings of the American Mathematical Society*, 145(7):3127–3136, 2017.

[22] Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, 2011.

[23] Steve Y Oudot. *Persistence theory: from quiver representations to data analysis*, volume 209. American Mathematical Society, 2015.

[24] Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. Number 2017-86. December 2017.

[25] Aaditya Ramdas, Nicolás Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[26] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 2015.

[27] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.

[28] The GUDHI Project. *GUDHI User and Reference Manual*. GUDHI Editorial Board, 2015.

[29] Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry*, 52(1):44–70, 2014.

[30] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[31] Jonathan Weed, Francis Bach, et al. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.