Imperial College London

A CONDITION FOR THE UNIQUENESS OF FRÉCHET MEANS OF PERSISTENCE DIAGRAMS

Yueqi Cao[1], Anthea Monod

Department of Mathematics, Imperial College London. **1–Funded by an EPSRC President's PhD Scholarship.**

MOTIVATION

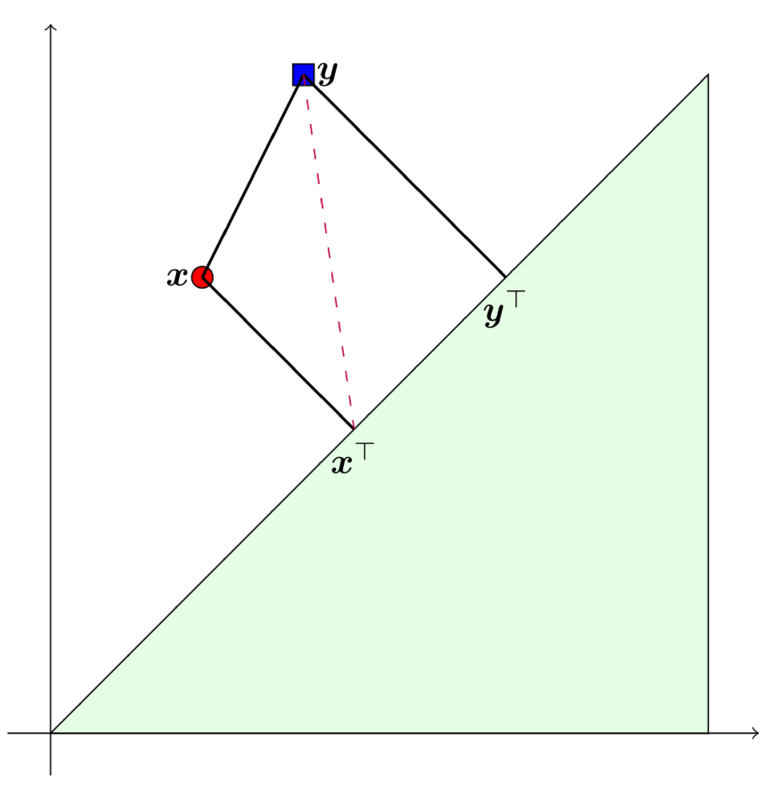
The Fréchet mean is an important statistical summary which has been defined and studied for sets of persistence diagrams (Mileyko, Mukherjee, and Harer, 2011; Munch et al., 2015; Turner et al., 2014). However, for a given set of persistence diagrams, the geometry of the space of persistence diagrams implies that the Fréchet mean may not be unique. This nonuniqueness prohibits convergence analysis for empirical Fréchet means of persistence diagrams (Cao and Monod, 2022), which motivates our study of what conditions yield a unique Fréchet mean and how fast do empirical Fréchet means converge to the population Fréchet mean.

SPACE OF PERSISTENCE DIAGRAMS

A persistence diagram is a locally finite multiset on the half plane $\Omega = \{(x, y) \in \mathbb{R}^2 \mid x < y\}$ together with points on the diagonal $\partial\Omega$ counted with infinite multiplicity. For any two persistence diagrams X and Y, define the 2-Wasserstein distance by

$$\mathrm{W}_2(X, Y) = \inf_{\phi} \left(\sum_{x \in X} \|x - \phi(x)\|^2 \right)^{\frac{1}{2}}$$

where $\phi : X \to Y$ ranges over all bijective matchings between X and Y. Let $\mathcal{D}_2$ be the set of persistence diagrams with finite total persistence. The metric space $(\mathcal{D}_2, \mathrm{W}_2)$ is known to be an Alexandrov space with nonnegative curvature (Turner et al., 2014). So for any two persistence diagrams in $\mathcal{D}_2$ there is a shortest path joining them, and for any triangle in $\mathcal{D}_2$, there is an isometric triangle in the Euclidean plane $\mathbb{R}^2$ with smaller angles.



Positive curvature at the boundary: Consider $X = \{x\}$ and $Y = \{y\}$ both with a single off-diagonal point and O with no off-diagonal point. The edges of the triangle $\triangle OXY \subseteq \mathcal{D}_2$ are plotted with solid lines. Let $\cdot^\top$ denote the projection to the diagonal. For the comparison triangle, given two edges $xx^\top, xy$, and the angle $\angle x^\top xy$, the length of the third edge is $\|x^\top y\|$. We see that $\|x^\top y\| > \|yy^\top\|$, indicating positive Alexsandrov curvature.

FRÉCHET MEANS AND GROUPINGS

Given a finite set of persistence diagrams $D_1, \ldots, D_L$, a *Fréchet mean* is any persistence diagram minimizing the following function

$$F(D) = \frac{1}{L} \sum_{i=1}^{L} \mathrm{W}_2^2(D, D_i)$$

Suppose D_i consists of k_i off-diagonal points for $i = 1, \ldots, L$. Set $K = \sum_{i=1}^{L} k_i$. A *grouping* is a $K \times L$ matrix G such that the ℓth column G^ℓ consists of k_ℓ off-diagonal points of D_ℓ and $K - k_\ell$ copies of the diagonal $\partial\Omega$. Let $\overline{G}_i$ denote the algorithmic mean of the ith row of G. The mean of a grouping $\mathrm{mean}(G)$ is a persistence diagram with points given by $\overline{G}_i$. The variance of a grouping G is defined as

$$\mathbb{V}(G) = \frac{1}{L} \sum_{\ell=1}^{L} \sum_{i=1}^{K} \|G_i^\ell - \overline{G}_i\|^2$$

Theorem (Munch et al., 2015; Turner et al., 2014) If $D_\star$ is a Fréchet mean of $D_1, \ldots, D_L$, then $D_\star = \mathrm{mean}(G_\star)$ for some grouping $G_\star$.

λ-SEPARATED GROUPINGS

For any grouping G, we have the following variance formula:

$$\mathbb{V}(G) = \frac{1}{L^2} \sum_{i=1}^{K} \sum_{1 \le w < \ell \le L} \|G_i^w - G_i^\ell\|^2 \qquad \text{(I)}$$

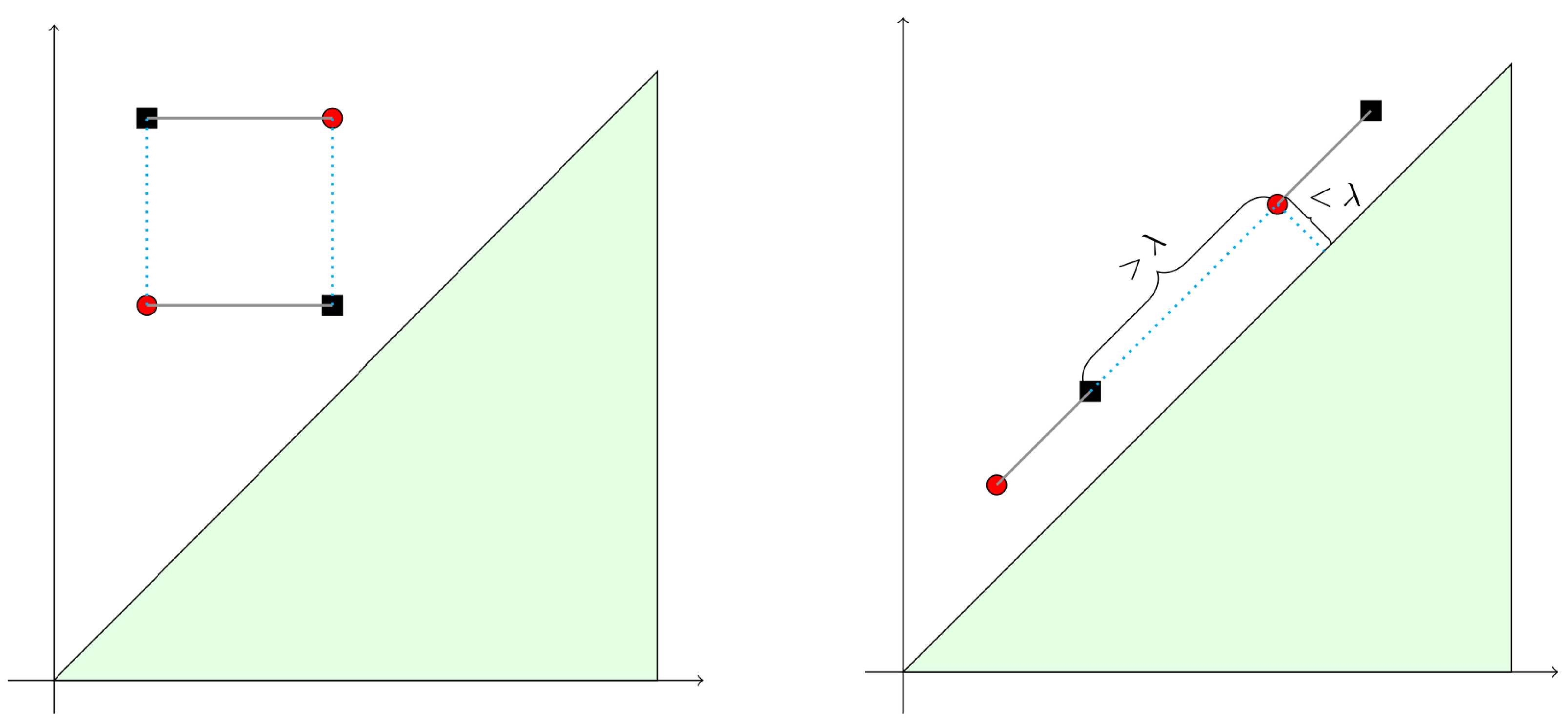
$$+ \sum_{i=1}^{K} \frac{L - s_i}{L^2 s_i} \sum_{1 \le w < \ell \le s_i} \|(G_i^{j_w})^\top - (G_i^{j_\ell})^\top\|^2 \qquad \text{(II)}$$

where $G_i^{j_\ell}, \ell = 1, \ldots, s_i$ ranges over all off-diagonal points in the ith row of G. The variance decomposes into two parts: (I) is the usual variance if all points are measured under Euclidean distance; (II) is the augmented variance due to curvature caused by the diagonal.

We say a grouping G *λ-separated*, if there exists $\lambda > 0$ such that

(a) For each row G_i, the diameter is bounded above by λ. i.e., $\|G_i^w - G_i^\ell\| < \lambda$ for all $w, \ell = 1, \ldots, L$;
(b) For two distinct rows G_i, G_j, the distance between G_i and G_j is bounded below by λ. i.e., $\|G_i^w - G_j^\ell\| > \lambda$ for all $w, \ell = 1, \cdots, L$;
(c) Off-diagonal points are bounded away from the diagonal by λ. i.e., $\|G_i^w - \partial\Omega\| > \lambda$ for $G_i^w \neq \partial\Omega$.

Theorem For $D_1, \ldots, D_L$, if there exists a λ-separated grouping G, then $\mathrm{mean}(G)$ gives the unique Fréchet mean.



Left: A grouping violates condition (a) and (b). The grouping is optimal but the Fréchet mean is not unique; **Right:** A grouping violates condition (c). The grouping is not optimal.

STATISTICAL PROPERTIES

Let $\rho = \frac{1}{L} \sum_{i=1}^{L} \delta_{D_i}$ be a discrete probability measure. If there exists a λ-separated grouping $\mathbf{G}$ for $D_1, \ldots, D_L$, then

- $D_\star = \mathrm{mean}(\mathbf{G})$ is the unique population Fréchet mean;
- $\widehat{D} = \mathrm{mean}(\widehat{G})$ is the unique empirical Fréchet mean for i.i.d. samples from ρ with empirical λ-separated grouping $\widehat{G}$;
- Let $\sigma^2 = \mathbb{V}(\mathbf{G})$. Then

$$\mathbb{E}[\mathrm{W}_2^2(\widehat{D}, D_\star)] \le \frac{\sigma^2}{B}$$

where B is the sample size.

REFERENCES

Cao, Yueqi and Anthea Monod (2022). "Approximating Persistent Homology for Large Datasets". In: *arXiv preprint arXiv:2204.09155*.

Mileyko, Yuriy, Sayan Mukherjee, and John Harer (2011). "Probability measures on the space of persistence diagrams". In: *Inverse Problems* 27.12, p. 124007.

Munch, Elizabeth et al. (2015). "Probabilistic Fréchet means for time varying persistence diagrams". In: *Electronic Journal of Statistics* 9.1, pp. 1173–1204.

Turner, Katharine et al. (2014). "Fréchet means for distributions of persistence diagrams". In: *Discrete & Computational Geometry* 52.1, pp. 44–70.