Provably Correct k-Means Clustering of Persistence Diagrams

Hajin Lee[1] **Kwangho Kim**[*1]

[1]Department of Statistics, Korea University

Abstract

We study k-means clustering of persistence diagrams through their Hilbert-space embeddings, focusing primarily on persistence landscapes. Although persistence landscapes are widely used in practice, it remains unclear when clustering them faithfully reflects clustering in the diagram space. We give simple, verifiable geometric conditions under which (i) nearest-center labels in the landscape space exactly agree with those in the diagram space, and (ii) the landscape k-means objective provably calibrates the diagram-space objective, leveraging tools from modern statistical learning theory. Combined with known fast rates for k-means in Hilbert spaces, our results show that landscape-based k-means provides a statistically and computationally efficient surrogate for diagram-space k-means with explicit performance guarantees and a controllable geometric distortion.

1 INTRODUCTION

Topological data analysis (TDA) provides a rigorous framework for representing the global geometric and topological structure of complex data, most notably via persistent homology and its summary in the form of persistence diagrams [Edelsbrunner and Harer, 2010, Chazal and Michel, 2021]. Because persistence diagrams are not directly amenable as inputs to standard machine learning algorithms, a common strategy is to replace each diagram by a more tractable functional or vector-valued summary and then apply off-the-shelf methods to these summaries. Examples include persistence landscapes [Bubenik, 2015], silhouettes [Chazal et al., 2014], kernel-based constructions such as persistence weighted Gaussian kernels [Kusano et al., 2016], and persistence images [Adams et al., 2017]. Empirically, such summaries have been shown to substantially improve the robustness of supervised learning methods, in particular deep learning architectures [e.g., Hofer et al., 2017, Carrière et al., 2020, Kim et al., 2020].

Nonetheless, systematic study of unsupervised learning, and clustering in particular, on the space of persistence diagrams remains relatively limited. Most existing approaches either treat diagram metrics as black-box distances and apply generic clustering to the resulting pairwise distance matrix [Gamble and Heo, 2010], or optimize directly in the diagram space via Fréchet means and Wasserstein barycenters [Turner et al., 2014]. Recent work includes scalable k-means in the space of persistence diagrams via optimal transport and entropic regularization [Lacombe et al., 2018], as well as diagram-space k-means variants that update cluster prototypes as Wasserstein/bottleneck barycenters [Cao et al., 2025].

However, clustering directly in the diagram space faces several intrinsic limitations. First, computing barycenters of persistence diagrams requires solving large-scale optimal transport problems, making each Lloyd update [Lloyd, 1982] substantially more expensive and numerically delicate than its Euclidean counterpart; this limits scalability even for moderate diagram sizes. Second, barycenters in the diagram space are often non-unique and potentially unstable, especially when diagrams contain many low-persistence points or when cluster assignments change slightly. Third, the space of persistence diagrams is not a Hilbert space, so standard convergence guarantees, variance decompositions, and fast-rate results for k-means do not apply. Finally, the lack of linear structure prevents the use of common acceleration techniques such as vector quantization, minibatch updates, or initialization schemes (e.g., k-means++ [Arthur and Vassilvitskii, 2006]), further complicating reliable large-scale clustering.

These limitations motivate the widespread use of functional or vector-valued summaries, as listed above, followed by standard clustering (e.g., k-means or spectral clustering) in

[*]Corresponding author.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:3323–3337.

the resulting feature space [Chan et al., 2022, Kim et al., 2018, Bhattacharya et al., 2025, Stolz et al., 2021, Reininghaus et al., 2015]. Yet, rigorous guarantees connecting clustering on such summaries back to clustering in the diagram space have remained scarce. In the absence of such guarantees, there is no principled control over how cluster assignments, risks, or decision boundaries in the summary space reflect the underlying geometry of persistence diagrams, leaving it unclear whether the resulting clusters correspond to genuine topological structures or artifacts of the chosen representation.

Our work addresses this gap by providing explicit, verifiable conditions under which k-means on Hilbert space embeddings is a sound surrogate for k-means in persistence-diagram space, both in terms of *risk* and *label preservation*. We place persistence landscape clustering within the statistical learning theory of vector quantization [Levrard, 2013, 2015], leveraging the stability of persistence landscapes to derive simple geometric conditions under which the surrogate is *certifiably faithful*. At the risk level, we show that the landscape k-means objective always lower-bounds the diagram-space objective and, under a restricted local bi-Lipschitz condition, the two risks are sandwiched within explicit multiplicative constants, yielding calibrated excess-risk bounds that transfer fast Hilbert-space k-means rates to the diagram space.

A further practical complication is that k-means centers in landscape space are arbitrary elements of L^p and need not correspond to valid persistence landscapes, hence to no actual persistence diagrams. We address this by post-processing each center via a simple "snapping" step: project it to the nearest sampled landscape, i.e., to the landscape of some observed diagram. We show that if the landscape-space cluster margins dominate the snapping error, then this projection leaves all sample labels unchanged and inflates the empirical landscape risk by at most a controlled additive term, which can be absorbed into the optimization error in our excess-risk bounds. Coupled with our calibration results, this yields diagram-space clusterings whose centers are genuine persistence diagrams, with explicit guarantees on label agreement and risk inflation.

Taken together, our results provide, to our knowledge, the first rigorous bridge between clustering in the persistence-diagram space and the widespread practical use of persistence landscapes, on which k-means can be implemented with standard Hilbert-space algorithms at substantially reduced computational costs. Geometrically, we derive finite-sample bi-Lipschitz and margin-type conditions under which Voronoi regions and nearest-center labels are preserved by the landscape embedding, together with a posteriori certification procedures. Statistically, we show that landscape k-means is an efficient surrogate for diagram-space k-means, with explicit control of geometric distortion, representation bias, and the effect of snapping to valid diagram centers. In this way, we both justify and precisely delineate when clustering persistence landscapes faithfully reflects structure in the underlying persistence diagrams. Our analysis extends directly to other stable Hilbert-valued summaries.

2 PRELIMINARIES

In this section, we briefly outline key TDA concepts and notations needed for subsequent development. For a more extensive introduction to TDA, we refer interested readers to Edelsbrunner and Harer [2010], Chazal and Michel [2021].

Simplex and Simplicial Complex. An *ℓ-simplex* σ_ℓ is the convex hull of $\ell+1$ affinely independent points in $\mathbb{R}^v$, i.e., $\sigma_\ell = \text{conv}\{u_0, \ldots, u_\ell\}$ for $u_i \in \mathbb{R}^v$ (e.g., 0-simplex: vertex, 1-simplex: edge, 2-simplex: triangle, etc.). η is a *face* of σ_ℓ if η is the convex hull of any non-empty subset of the $\ell+1$ vertices of σ_ℓ. A *simplicial complex* W is a finite collection of simplices such that (i) any face of a simplex in W is also in W, and (ii) the intersection of any two simplices in W is either empty or a common face of both.

Persistent Homology and Diagrams. A *filtration* of a simplicial complex W is a collection of nested subcomplexes $\mathcal{F} = \{W(t) \subseteq W\}_{t\in\mathbb{R}}$ such that $t_1 \le t_2$ implies $W(t_1) \subseteq W(t_2)$. Filtrations are often constructed by specifying a monotonic *filtration function* $f : W \to \mathbb{R}$ that satisfies $f(\eta) \le f(\sigma)$ whenever η is a face of σ. By setting $W(t) := f^{-1}((-\infty, t])$, we immediately obtain a nested family of subcomplexes. Given a filtration $\mathcal{F}$, *persistent homology* tracks the birth and death of homological features; a homological feature is said to be born at a and to die at b if it appears in $W(a)$ and disappears in $W(b)$, where $a < b$. A *persistence diagram* $D(\mathcal{F})$ is a finite multiset of all such birth-death pairs $(a, b) \in (\mathbb{R} \cup \{\infty\})^2$, and we let $D^d(\mathcal{F})$ denote the d-th persistence diagram corresponding to d-dimensional homological features (e.g., $d = 0$: connected components, $d = 1$: loops, $d = 2$: cavity, etc.). For notational simplicity, the filtration $\mathcal{F}$ will be omitted when understood from context. Throughout this paper, we assume that all persistence diagrams are bounded and consist of finitely many points (a, b) such that $-\infty < a < b < \infty$.

Bottleneck Distance. The space of persistence diagrams can be endowed with a metric structure, most notably the bottleneck distance. Let $Diag := \{(t, t) \,|\, t \in \mathbb{R}\} \subset \mathbb{R}^2$ represent the diagonal subset with infinite multiplicity. A *matching* between two persistence diagrams D and D' is a subset $m \subset (D \cup Diag) \times (D' \cup Diag)$ such that every point in D and D' appears only once in m. Then, the *bottleneck distance* between two persistence diagrams D and D' is given by

$$d_B(D, D') = \inf_{\text{matching } m} \max_{(p,q)\in m} \|p - q\|_\infty.$$

Persistence Landscapes. Persistence diagrams are intrinsically multisets, which complicates their use as inputs to machine learning pipelines. To address this, one may consider alternative representations such as persistence landscapes [Bubenik, 2015], which are mappings of persistence diagrams into a functional Hilbert space. Given a persistence diagram D, we first define, for each point $p = (a, b) \in D$, a piecewise linear tent function

$$\mathrm{tri}_p(t) = \max\{0,\, \min\{t - a, b - t\}\}.$$

Then, the *persistence landscape* λ of D is a sequence of functions $\lambda = \{\lambda_j\}_{j\in\mathbb{N}}$, where each λ_j is defined as

$$\lambda_j(t) = \mathrm{jmax}_{p\in D}\, \mathrm{tri}_p(t), \quad j \in \mathbb{N},\ t \in \mathbb{R}.$$

jmax is the j-th largest value in the set, and λ_j is called the j-th persistence landscape.

A key property of persistence landscapes is their stability with respect to the supremum norm [Bubenik, 2015, Theorem 13]. Let λ_j and λ'_j be the j-th persistence landscapes of persistence diagrams D and D', respectively. Then, for all $j \in \mathbb{N}$,

$$\|\lambda_j - \lambda'_j\|_\infty \le d_B(D, D').$$

Furthermore, for any $J \in \mathbb{N}$, if we let $\lambda_{1:J} \coloneqq \{\lambda_j\}_{j=1}^J$ and define $\|\lambda_{1:J}\|_\infty \coloneqq \frac{1}{J}\sum_{j=1}^J \|\lambda_j\|_\infty$, it is immediately apparent from the above inequality that

$$\|\lambda_{1:J} - \lambda'_{1:J}\|_\infty \le d_B(D, D'). \tag{1}$$

The stability results described above are instrumental in developing our results in subsequent sections.

3 SETUP AND NOTATIONS

For the remainder of this paper, we assume a fixed homology dimension d and suppress all notational dependence, as the subsequent results hold for an arbitrary choice of d. Let $(\mathcal{D}, d_B)$ be the metric space of persistence diagrams equipped with the bottleneck distance. We treat persistence landscapes as functions that reside in a separable Hilbert space equipped with the $\|\cdot\|_p$ norm ($1 \le p \le \infty$) [Bubenik, 2015], and define the j-th landscape map $\Lambda_j : \mathcal{D} \to L^p$. The K unique cluster centers of diagrams and j-th landscapes are denoted as $C = \{c_1, \ldots, c_K\} \subset \mathcal{D}$ and $M_j = \{\mu_1^j, \ldots, \mu_K^j\} \subset L^p$, respectively. In this work, we restrict our attention to the top-J landscapes, as higher order landscapes typically capture noisy fluctuations that are not representative of the global structure. We write $\Lambda_{1:J}(D) = \{\Lambda_j(D)\}_{j=1}^J$ to represent the top-J landscapes of D, and we let $M_{1:J} = \cup_{j=1}^J M_j$ be the associated landscape cluster centers. For $[K] \coloneqq \{1, \ldots, K\}$ and a fixed cluster $k \in [K]$, we set $\mu_k^{1:J} = \{\mu_k^1, \ldots, \mu_k^J\}$ as the top-J centroids corresponding to cluster k. We note that the landscape centroids $M_{1:J}$ consist of arbitrary elements in L^p that need not correspond to valid landscapes, a subtlety further elaborated in Section 4. Finally, $S = \{D_i\}_{i=1}^n \subset \mathcal{D}$ denotes a finite sample of persistence diagrams. For convenience of presentation, we adopt the normalized form $\|\lambda_{1:J}\|_p \coloneqq \frac{1}{J}\sum_{j=1}^J \|\lambda_j\|_p$ for distances, rather than the classical definition $\|\lambda_{1:J}\|_p = \sum_{j=1}^J \|\lambda_j\|_p$. Hereafter, we implicitly assume the use of top-J landscapes and drop the notation $1{:}J$ for simplicity.

Nearest Cluster Label. For any $D_i \in S$, we define the *nearest-center indices* for diagrams and landscapes as

$$\pi(D_i \,|\, C) \in \arg\min_{k\in[K]} d_B(D_i, c_k),$$
$$\pi(\Lambda(D_i) \,|\, M) \in \arg\min_{k\in[K]} \|\Lambda(D_i) - \mu_k\|_\infty,$$

respectively, with a fixed deterministic tie-breaking rule (e.g., smallest index).

Margins. Assume that $d_B(D_i, c_{\pi(D_i|C)}) \neq 0$, and let

$$r_{ik} \coloneqq \frac{d_B(D_i, c_k)}{d_B(D_i, c_{\pi(D_i|C)})}. \tag{2}$$

Then, the *diagram-space ratio margin* is defined as

$$r_i \coloneqq \min_{k\neq\pi(D_i|C)} r_{ik} \ \ge 1.$$

This is the ratio between the distances from D_i to its two nearest diagram centers. Next, let

$$\Delta_{ik} \coloneqq \|\Lambda(D_i) - \mu_k\|_\infty - \|\Lambda(D_i) - \mu_{\pi(\Lambda(D_i)|M)}\|_\infty. \tag{3}$$

Then, the *landscape-space margin* is defined as

$$\Delta_i \coloneqq \min_{k\neq\pi(\Lambda(D_i)|M)} \Delta_{ik} \ \ge 0,$$

which represents the gap between the distances from $\Lambda(D_i)$ to its two nearest landscape centers.

Risk. Let $\rho : \mathbb{R}_+ \to \mathbb{R}_+$ be a nondecreasing function and $S = \{D_i\}_{i=1}^n$ a finite sample. Given a diagram centroid $C = \{c_1, \ldots, c_K\}$ and a landscape centroid $M = \{\mu_1, \ldots, \mu_K\}$, define the diagram and landscape *empirical risks* as

$$\widehat{R}_{\mathcal{D}}(C) \coloneqq \frac{1}{n}\sum_{i=1}^n \rho\Big(\min_{k\in[K]} d_B(D_i, c_k)\Big),$$
$$\widehat{R}_{\Lambda}(M) \coloneqq \frac{1}{n}\sum_{i=1}^n \rho\Big(\min_{k\in[K]} \|\Lambda(D_i) - \mu_k\|_\infty\Big),$$

respectively. Moreover, if D is a random persistence diagram independently distributed with law P on $\mathcal{D}$, the diagram and landscape *population risks* are given as

$$R_{\mathcal{D}}(C) \coloneqq \mathbb{E}_P\Big[\rho\Big(\min_{k\in[K]} d_B(D, c_k)\Big)\Big],$$
$$R_{\Lambda}(M) \coloneqq \mathbb{E}_P\Big[\rho\Big(\min_{k\in[K]} \|\Lambda(D) - \mu_k\|_\infty\Big)\Big],$$

respectively. Throughout, we assume that all diagram samples and centroids lie within a d_B-bounded support in $\mathcal{D}$.

Optimal Cluster Centers. The *population-optimal cluster centers* for diagrams and landscapes are

$$C^* \in \arg\min_{C \in \mathcal{C}_K} R_{\mathcal{D}}(C), \qquad M^* \in \arg\min_{M \in \mathcal{M}_K} R_{\Lambda}(M),$$

where $\mathcal{C}_K \subset \text{supp}(P)$ and $\mathcal{M}_K \subset L^p$ denote the set of all diagram and landscape centroids with K clusters, respectively. We also define a third optimal cluster center

$$C^*_{\Lambda} \in \arg\min_{C \in \mathcal{C}_K} R_{\Lambda}(\Lambda(C)),$$

which is the diagram-space cluster center whose landscape mapping minimizes the population landscape risk.

4 THEORETICAL PROPERTIES

In this section, we establish the theoretical foundation for landscape-based surrogate clustering of persistence diagrams. To guide readers through the theoretical development, we begin by providing an overview of the main results and outlining how they fit together. In Section 4.1, we characterize when the cluster assignments obtained in the landscape space remain consistent with those in the diagram space. Given that landscape centroids are not guaranteed to be valid persistence landscapes corresponding to diagram centroids, we establish conditions under which the diagram centroids C, their landscape representations $\Lambda(C)$, and the landscape centroids M all induce identical cluster labels (Theorems 4.1 and 4.2), and introduce a simple "snapping" procedure to recover label-preserving diagram centroids (Corollary 4.3). We proceed to establish risk calibration results in Section 4.2, analyzing how the landscape objective controls the diagram-space objective. Here, we restrict our attention to landscape centroids lying in the image $\Lambda(\mathcal{D})$, which enables us to relate the landscape risk to the diagram risk through the stability relation in (1) together with a local bi-Lipschitz assumption. We first show that, for any diagram centroid, the landscape risk lower bounds the diagram risk (Theorem 4.4), and that the diagram-space excess risk is controlled by its corresponding landscape excess risk (Theorem 4.5). We then specialize the analysis of the diagram-space excess risk to the specific case where an approximate empirical minimizer is used as the diagram centroid (Theorem 4.6), and examine how the landscape excess risk convergence rates are transferred to the diagram excess risk convergence rates (Corollary 4.7 and 4.7). Finally, in Section 4.3, we connect these theoretical results to the practical k-means algorithm by quantifying the effect of snapping on empirical near-optimality (Lemma 4.9), and we present the resulting algorithmic procedure in Section 4.4.

4.1 LABEL PRESERVATION

In standard k-means, centroids are allowed to be arbitrary points in the ambient feature space. Thus, the landscape cluster centers $M = \{\mu_1, \ldots, \mu_K\}$ need not belong to the image $\Lambda(\mathcal{D})$ of persistence diagrams, but are rather arbitrary elements in L^p. For instance, the resulting cluster centers of the standard Lloyd's k-means algorithm [Lloyd, 1982] are the averages of all landscapes assigned to each cluster, which generally does not yield a valid persistence landscape. Thus, the landscape centers M may not translate directly to the diagram-space cluster centers, complicating interpretation and analysis. To remedy this, we show that, under suitable assumptions, there exists a centroid C in the space of persistence diagrams that preserves the cluster labels induced by M. We first state the necessary assumption required to establish our results.

Assumption 1 (Margin conditions)**.** *For all $D_i \in S$, the diagram-space ratio margin r_i is strictly larger than 1, and the landscape-space margin Δ_i is strictly larger than 0, i.e., there exist constants r_0 and ϵ_0 such that*

$$r_i = \min_{k \neq \pi(D_i|C)} r_{ik} > r_0 \geq 1,$$
$$\Delta_i = \min_{k \neq \pi(\Lambda(D_i)|M)} \Delta_{ik} > \epsilon_0 \geq 0.$$

This condition ensures the uniqueness of nearest cluster in both the diagram and landscape space for all samples.

Assumption 1 enforces a natural cluster separation in both the diagram and landscape spaces and is routinely adopted in the k-means and vector-quantization literature (see, for example, Levrard [2013, 2015], Ostrovsky et al. [2013], Awasthi and Sheffet [2012]).

Under Assumption 1, Theorem 4.1 shows that for a given landscape centroid M, there exists a corresponding diagram centroid C that preserves the cluster labels, i.e, diagram-space clustering using C yields labels identical to those induced by M in the landscape space.

Theorem 4.1 (Label-preserving diagram centroid)**.** *Let $S = \{D_i\}_{i=1}^n$ be a finite sample, $M = \{\mu_1, \ldots, \mu_K\}$ a landscape centroid, and $C = \{c_1, \ldots, c_K\}$ a diagram centroid. Suppose there exists an $\alpha_n := \alpha_n(S, C) \in (0, 1]$ such that for all $D_i \in S$,*

$$\alpha_n \leq \frac{\|\Lambda(D_i) - \Lambda(c_{\pi(\Lambda(D_i)|\Lambda(C))})\|_\infty}{d_B(D_i, c_{\pi(\Lambda(D_i)|\Lambda(C))})},$$

where $d_B(D_i, c_{\pi(\Lambda(D_i)|\Lambda(C))}) \neq 0$. For each $k \in [K]$, define $\tau_k := \|\Lambda(c_k) - \mu_k\|_\infty$, and let $\tau := \max_{k \in [K]} \tau_k$. Assume further that Assumption 1 holds with the margin constants

1. *$r_0 = 1/\alpha_n$, and*

2. $\epsilon_0 = 2\tau$.

Then, C is label-preserving with respect to M, in the sense that for every $D_i \in S$,

$$\pi(D_i \,|\, C) \;=\; \pi(\Lambda(D_i) \,|\, \Lambda(C)) \;=\; \pi(\Lambda(D_i) \,|\, M).$$

The diagram-space ratio margin condition with $r_0 = 1/\alpha_n$ ensures that $\pi(D_i \,|\, C) = \pi(\Lambda(D_i) \,|\, \Lambda(C))$, while the landscape-space margin condition with $\epsilon_0 = 2\tau$ guarantees that $\pi(\Lambda(D_i) \,|\, \Lambda(C)) = \pi(\Lambda(D_i) \,|\, M)$. Furthermore, the existence of the constant α_n can be verified a posteriori by taking minimum ratio over all diagrams in S.

While the margin conditions in Theorem 4.1 provide a clean and convenient criterion for ensuring cluster label preservation, they are *uniform* sufficient conditions expressed through global constants over all $D_i \in S$ and $k \in [K]$, thereby reflecting a worst-case guarantee. Consequently, these margin conditions may be overly conservative, requiring a strong separation between cluster centers. Theorem 4.2 relaxes these uniform requirements by introducing pointwise conditions that hold separately for each sample D_i and each competing cluster k.

Theorem 4.2 (Pointwise label preservation)**.** *For a landscape centroid $M = \{\mu_1, \ldots, \mu_K\}$ and a diagram centroid $C = \{c_1, \ldots, c_K\}$, let $\tau_k = \|\Lambda(c_k) - \mu_k\|_\infty$ for $k \in [K]$, as in Theorem 4.1. Given a finite sample $S = \{D_i\}_{i=1}^n$, let us define*

$$\alpha_{ik} := \frac{\|\Lambda(D_i) - \Lambda(c_k)\|_\infty}{d_B(D_i, c_k)} \in (0, 1], \quad i \in [n],\ k \in [K],$$

where $d_B(D_i, c_k) \neq 0$ for all k. If, for every $i \in [n]$,

1. $r_{ik} > 1/\alpha_{ik}$, *for all $k \neq \pi(D_i \,|\, C)$, and*
2. $\Delta_{ik} > \tau_{\pi(\Lambda(D_i)|M)} + \tau_k$, *for all $k \neq \pi(\Lambda(D_i) \,|\, M)$,*

where r_{ik} and Δ_{ik} are defined as in (2) *and* (3)*, then,*

$$\pi(D_i \,|\, C) \;=\; \pi(\Lambda(D_i) \,|\, \Lambda(C)) \;=\; \pi(\Lambda(D_i) \,|\, M),$$

for all $D_i \in S$.

Theorem 4.2 replaces the global constants in Theorem 4.1 with margins that are specific to each sample and competing cluster, thereby yielding less restrictive conditions for label preservation. As before, the first condition ensures that $\pi(D_i \,|\, C) = \pi(\Lambda(D_i) \,|\, \Lambda(C))$, while the second condition guarantees $\pi(\Lambda(D_i) \,|\, \Lambda(C)) \;=\; \pi(\Lambda(D_i) \,|\, M)$. Moreover, the existence of α_{ik} in Theorem 4.2 is empirically verifiable by evaluating the minimal slack $\kappa := \min_i \min_{k \neq \pi(D_i|C)} \{\alpha_{ik}\, d_B(D_i, c_k) - d_B(D_i, c_{\pi(D_i|C)})\}$. If κ is positive, the condition $d_B(D_i, c_{\pi(D_i|C)}) < \alpha_{ik}\, d_B(D_i, c_k)$ for every $k \neq \pi(D_i \,|\, C)$ is satisfied for all $D_i \in S$.

In practice, it may be non-trivial to identify a label-preserving diagram centroid C described in Theorems 4.1 and 4.2. A useful heuristic is to assign the data point D_i whose landscape mapping lies closest to μ_k as the diagram cluster center c_k. Provided that such data points satisfy the conditions specified in Theorem 4.2, the resulting diagram centroid will inherit the label-preserving property, as formally stated in Corollary 4.3. We further note that the pointwise conditions in Theorem 4.2 imply the uniform conditions in Theorem 4.1.

Corollary 4.3 (Landscape centroid snapping)**.** *Consider a finite sample $S = \{D_i\}_{i=1}^n$. Given a landscape centroid $M = \{\mu_1, \ldots, \mu_K\}$, we assign for each $k \in [K]$,*

$$c_k^{\mathrm{snap}} = \arg\min_{D_i \in S} \|\Lambda(D_i) - \mu_k\|_\infty,$$

and define $C^{\mathrm{snap}} := \{c_1^{\mathrm{snap}}, \ldots, c_K^{\mathrm{snap}}\} \subset S$. If C^{snap} satisfies the pointwise margin conditions in Theorem 4.2, then C^{snap} retains the label-preserving property, i.e.,

$$\pi(D_i \,|\, C^{\mathrm{snap}}) = \pi(\Lambda(D_i) \,|\, \Lambda(C^{\mathrm{snap}})) = \pi(\Lambda(D_i) \,|\, M).$$

The above procedure is referred to as landscape centroid *snapping*, as we "snap" the landscape centroids to a nearby sampled persistence landscape $\Lambda(D_i)$.

An important observation in Theorems 4.1 and 4.2 is that, given a label-preserving diagram centroid C, the cluster labels are not affected by replacing the landscape centroid M with a landscape-mapped diagram centroid $\Lambda(C)$, i.e., $\pi(\Lambda(D_i) \,|\, \Lambda(C)) \;=\; \pi(\Lambda(D_i) \,|\, M)$. Whereas understanding the properties of the diagram space through M is challenging, as M may not translate into the diagram space, working with $\Lambda(C)$ instead makes this analysis considerably more tractable. To this end, Theorems 4.1 and 4.2 establish a valuable connection that enables us to assess diagram-space risks by leveraging the relationship between C and $\Lambda(C)$, as discussed in the following section.

4.2 CONTROLLING DIAGRAM-SPACE RISK

In this section, we analyze how clustering with centroids $\Lambda(C)$ in the landscape space translates into corresponding risk behavior in the diagram space. Imposing the structural constraint $M = \Lambda(C)$ on the landscape centroids, we first derive a lower bound on the diagram risk via the stability result in (1).

Theorem 4.4 (Surrogate risk domination)**.** *For a finite sample $S = \{D_i\}_{i=1}^n$ and any $C \in \mathcal{C}_K$, it holds that*

$$\widehat{R}_\Lambda(\Lambda(C)) \;\leq\; \widehat{R}_{\mathcal{D}}(C), \qquad R_\Lambda(\Lambda(C)) \;\leq\; R_{\mathcal{D}}(C).$$

Theorem 4.4 provides a general lower bound on both the empirical and population diagram risks in terms of their

corresponding landscape risks. However, for landscapes to serve as a reliable surrogate for clustering in the diagram space, we require an upper bound ensuring that minimization of the landscape risk leads to corresponding reduction in the diagram risk. Accordingly, we introduce a condition which enables *population-level* control of an upper bound on the diagram risk.

Assumption 2 (Local bi-Lipschitzness of Λ)**.** *There exists an $\alpha \in (0, 1]$ such that for all $D \in \mathrm{supp}(P)$,*

$$\begin{aligned}\alpha d_B(D, c_{\pi(\Lambda(D)|\Lambda(C))}) \;&\le\; \|\Lambda(D) - \Lambda(c_{\pi(\Lambda(D)|\Lambda(C))})\|_\infty \\ &\le\; d_B(D, c_{\pi(\Lambda(D)|\Lambda(C))}).\end{aligned}$$

This condition enables a local bi-Lipschitz embedding between each diagram cluster center and diagrams in its landscape-space neighborhood, i.e., diagrams whose landscape images are closest to that of the center. It ensures that, locally, separation between diagrams is not collapsed under the landscape map, as any diagram-space margin is preserved up to the factor α in the landscape space.

Assumption 2 is a *local* property that remains consistent with known impossibility results that preclude global bi-Lipschitz embedding of the entire diagram space into a Hilbert space. Under this assumption, we can derive explicit upper bounds on both the diagram population risk and the diagram-space excess risk, as presented in Theorem 4.5.

Theorem 4.5 (Calibrated population surrogate)**.** *Let the loss ρ be a power function $\rho(t) = t^q$ with $q \ge 1$, and assume that Assumption 2 holds. Then, for any $C \in \mathcal{C}_K$, we have*

$$R_{\mathcal{D}}(C) \;\le\; \alpha^{-q} R_\Lambda(\Lambda(C)),$$

and

$$\begin{aligned}R_{\mathcal{D}}(C) - R_{\mathcal{D}}(C^*) \;&\le \alpha^{-q}\{R_\Lambda(\Lambda(C)) - R_\Lambda(\Lambda(C^*_\Lambda))\} \\ &+\; (\alpha^{-q} - 1)\, R_\Lambda(\Lambda(C^*_\Lambda)),\end{aligned}$$

where C^* and C^*_Λ are the optimal diagram centroids minimizing the diagram population risk $R_{\mathcal{D}}(C)$ and the landscape population risk $R_\Lambda(\Lambda(C))$, respectively.

Thus, the diagram-space excess risk $R_{\mathcal{D}}(C) - R_{\mathcal{D}}(C^*)$ is controlled (up to a multiplicative constant and a fixed offset) by its corresponding landscape excess risk.

We next turn our attention to establishing a more precise characterization of how landscape-based clustering optimization manifests in terms of diagram-space excess risk. Namely, we examine the diagram-space excess risk induced by $\widehat{C}^n_\Lambda \in \mathcal{C}_K$, which is the diagram centroid that approximately minimizes the landscape empirical risk $\widehat{R}_\Lambda(\Lambda(C))$.

Theorem 4.6 (Diagram-space excess risk via empirical surrogate clustering)**.** *Let Assumption 2 hold with $\rho(t) = t^q$, $q \ge 1$. For a sample $D_1, \ldots, D_n \overset{i.i.d.}{\sim} P$, let $\widehat{C}^n_\Lambda \in \mathcal{C}_K$ be an approximate empirical minimizer,*

$$\widehat{R}_\Lambda(\Lambda(\widehat{C}^n_\Lambda)) \;\le\; \inf_{C \in \mathcal{C}_K} \widehat{R}_\Lambda(\Lambda(C)) \;+\; \varepsilon_n, \tag{4}$$

for some optimization error $\varepsilon_n \ge 0$. Suppose furthermore that there exists a sequence $\Gamma_n \to 0$ such that, with probability at least $1 - \delta_n$,

$$\sup_{C \in \mathcal{C}_K} \left|R_\Lambda(\Lambda(C)) - \widehat{R}_\Lambda(\Lambda(C))\right| \;\le\; \Gamma_n. \tag{5}$$

Then on the event (5)*, the diagram-space excess risk of $\widehat{C}^n_\Lambda$ satisfies*

$$\begin{aligned}R_{\mathcal{D}}(\widehat{C}^n_\Lambda) - R_{\mathcal{D}}(C^*) \;&\le \alpha^{-q}\big(\varepsilon_n + 2\Gamma_n\big) \\ &\quad + (\alpha^{-q} - 1)\, R_\Lambda(\Lambda(C^*_\Lambda)).\end{aligned} \tag{6}$$

with probability at least $1 - \delta_n$. In particular, if $\varepsilon_n \to 0$, $\delta_n \to 0$, and $\Gamma_n \to 0$ as $n \to \infty$, then

$$R_{\mathcal{D}}(\widehat{C}^n_\Lambda) \;\longrightarrow\; R_{\mathcal{D}}(C^*) \quad \textit{in probability,}$$

up to the irreducible structural offset $(\alpha^{-q} - 1)\, R_\Lambda(\Lambda(C^*_\Lambda))$.

Theorem 4.6 presents a principled analysis of the diagram-space excess risk given an approximate empirical minimizer $\widehat{C}^n_\Lambda$, obtained via optimization in the Hilbert space. The diagram-space excess risk of $\widehat{C}^n_\Lambda$ exhibits an upper bound that depends on $(\alpha^{-q} - 1)\, R_\Lambda(\Lambda(C^*_\Lambda))$, ε_n, and Γ_n. The first term is an irreducible bias that inherently depends on the complexity of the given problem, and is minimized when α is close to 1, i.e., when the diagram margins are well preserved in the landscape space. ε_n is the optimization error that depends on the optimization method, and Γ_n is a uniform general bound that bounds the discrepancy between the landscape population risk and empirical risk for all landscape centroids $\Lambda(C)$ with probability at least $1 - \delta_n$.

Once the landscape-space excess risk for $\Lambda(\widehat{C}^n_\Lambda)$ admits a known convergence rate, we can extend the analysis beyond upper bounds to derive the convergence rate of the diagram-space excess risk.

Corollary 4.7 (Inheritance of Hilbert space rates)**.** *Assume the setting of Theorem 4.6. If there exists a rate $\varphi_n \to 0$ such that $R_\Lambda(\Lambda(\widehat{C}^n_\Lambda)) - R_\Lambda(\Lambda(C^*_\Lambda)) = O_p(\varphi_n)$, then,*

$$R_{\mathcal{D}}(\widehat{C}^n_\Lambda) - R_{\mathcal{D}}(C^*) \;=\; O_p\big(\alpha^{-q}\varphi_n\big) + (\alpha^{-q} - 1)\, R_\Lambda(\Lambda(C^*_\Lambda)).$$

Thus, the diagram-space excess risk of $\widehat{C}^n_\Lambda$ inherits the Hilbert space convergence rate up to the factor α^{-q} plus the fixed structural problem-dependent bias $(\alpha^{-q} - 1)\, R_\Lambda(\Lambda(C^*_\Lambda))$. Accordingly, existing results for k-means convergence in general Hilbert spaces can be leveraged to further characterize the convergence rates of diagram excess risk, as detailed in Corollary 4.8.

Corollary 4.8 (Ideal surrogate theorem under fast rates)**.** *Assume the setting of Theorem 4.6 with $\rho(t) = t^2$ (standard k-means). Suppose further that:*

1. *$\widehat{C}_\Lambda^n$ is an ε_n-approximate empirical minimizer of $\widehat{R}_\Lambda$ as in Theorem 4.6.*
2. *The uniform generalization bound* (5) *holds with $\Gamma_n = O_p(\varphi_n)$.*
3. *Under a suitable margin/regularity condition for k-means in the Hilbert space, we have a fast rate*

$$R_\Lambda(\Lambda(\widehat{C}_\Lambda^n)) - R_\Lambda(\Lambda(C_\Lambda^*)) = O_p(\varphi_n),$$

where $\varphi_n = n^{-1/2}$ or $\varphi_n = n^{-1}$.

Then, under the first and third conditions, the diagram-space excess risk satisfies

$$R_{\mathcal{D}}(\widehat{C}_\Lambda^n) - R_{\mathcal{D}}(C^*) \;=\; O_p\big(\alpha^{-2}\varphi_n\big) + (\alpha^{-2}-1)\, R_\Lambda(\Lambda(C_\Lambda^*)),$$

where $(\alpha^{-2}-1)\, R_\Lambda(\Lambda(C_\Lambda^*)) \geq 0$ is the fixed problem-dependent bias defined in Theorem 4.6.

Under the first and second conditions, any ε_n-approximate minimizer of the empirical landscape k-means objective yields

$$\begin{aligned} R_{\mathcal{D}}(\widehat{C}_\Lambda^n) - R_{\mathcal{D}}(C^*) \;&= O_p\big(\alpha^{-2}\varepsilon_n\big) + O_p\big(\alpha^{-2}\varphi_n\big) \\ &\quad + (\alpha^{-2}-1)\, R_\Lambda(\Lambda(C_\Lambda^*)). \end{aligned}$$

That is, diagram-space excess risk of order $O(\varepsilon_n)$ plus a statistical term of order $O_p(\varphi_n)$, both up to the factor α^{-2}, with an additional fixed structural bias term.

The second condition $\Gamma_n = O_p(\varphi_n)$ is conceptually standard and typically follows from classical empirical-process or Rademacher complexity arguments; it mainly requires that the landscape risk class indexed by $\Lambda(C)$ is not too rich. By contrast, the third condition directly imposes a fast-rate excess-risk bound for $R_\Lambda(\Lambda(\widehat{C}_\Lambda^n)) - R_\Lambda(\Lambda(C_\Lambda^*))$, which typically requires stronger margin-type assumptions. The achievable rate φ_n depends on the strength of these additional, though still commonly used in Hilbert-space k-means analyses, regularity conditions. Under the mild assumption that the D_i are i.i.d. and supported on a d_B-bounded subset of $\mathcal{D}$, one obtains a convergence rate of $O_p(n^{-1/2})$ [Biau et al., 2008]. When further separation or curvature conditions are imposed, such as the margin condition in Levrard [2015, Definition 2.1], the rate improves to $O_p(n^{-1})$. Thus, clustering in the Hilbert space, rather than in the diagram space, allows us to attain the fast Hilbert-space rate guarantees for the diagram excess risk.

4.3 BRIDGING $\Lambda(C)$ AND M

Thus far, our analysis has been devoted to assessing the diagram-space excess risk induced by a landscape centroid constrained to the form $\Lambda(C)$. In practice, however, landscape centroids returned by the k-means algorithm rarely lie in the image $\Lambda(\mathcal{D})$ of persistence diagrams, as each $\mu_k \in M$ is an arbitrary element in L^p. To address this, we have proposed a data-driven heuristic in Corollary 4.3 that "snaps" each μ_k to a nearby sampled landscape $\Lambda(D_i)$ while preserving the cluster labels. Thus, our interest now lies in quantifying how the discrepancy between the k-means output landscape centroid M and the snapped landscape centroid $\Lambda(C^{\mathrm{snap}})$ influences the results presented in previous sections. Accordingly, Lemma 4.9 establishes an empirical near-optimality result for the diagram centroid C^{snap} obtained by snapping the k-means output landscape centers.

Lemma 4.9 (Empirical near-optimality of snapped centers)**.** *Assume $\rho(t) = t^q$ with $q \geq 1$. Let $S = \{D_i\}_{i=1}^n$ be a finite sample, and suppose $\widehat{M}^n = \{\mu_1, \ldots, \mu_K\}$ is the landscape centroid (not necessarily valid landscapes) returned by a k-means procedure in L^p. Assume that, for some optimization error $\gamma_n \geq 0$,*

$$\widehat{R}_\Lambda(\widehat{M}^n) \;\leq\; \inf_{M \in \mathcal{M}_K} \widehat{R}_\Lambda(M) \;+\; \gamma_n.$$

Let $C^{\mathrm{snap}} = \{c_1^{\mathrm{snap}}, \ldots, c_K^{\mathrm{snap}}\} \subset S$ be the snapped centers as in Corollary 4.3, such that $\|\Lambda(c_k) - \mu_k\|_\infty \leq \tau_k$ for $k \in [K]$ with $\tau := \max_{k\in[K]} \tau_k$. Define the maximal data–center distance in landscape space by

$$B := \max_{i\in[n]} \left\{ \min_{k\in[K]} \|\Lambda(D_i) - \mu_k\|_\infty, \min_{k\in[K]} \|\Lambda(D_i) - \Lambda(c_k)\|_\infty \right\}.$$

$L_q := qB^{q-1}$ is the Lipschitz constant of $\rho(t) = t^q$ such that $|x^q - y^q| \leq L_q |x - y|$ for all $x, y \in [0, B]$. Then,

$$\widehat{R}_\Lambda(\Lambda(C^{\mathrm{snap}})) = \inf_{C \in \mathcal{C}_K} \widehat{R}_\Lambda(\Lambda(C)) + \varepsilon_n, \quad \varepsilon_n := L_q\, \tau + \gamma_n.$$

Given this choice of ε_n, the snapped diagram centroid C^{snap} satisfies the empirical near-optimality condition (4) characterizing an approximate empirical minimizer. Hence, the diagram-space excess-risk conclusions of Theorem 4.6 hold for C^{snap}. In particular, if $\tau \to 0$ and $\gamma_n \to 0$, then $\varepsilon_n \to 0$. The snapping bound τ is a fully data-dependent term that becomes 0 in the special case that M lies in the image $\Lambda(\mathcal{D})$ of persistence diagrams, and γ_n is the Hilbert space optimization error that depends on the optimization algorithm.

4.4 ALGORITHMIC PROCEDURE

It is important to note that the analysis in previous sections were independent of the optimization algorithm. We now describe an algorithm for performing k-means clustering on persistence landscapes. A profound advantage of clustering in the Hilbert space, in contrast to the diagram space, is that it permits standard operations such as averaging, thereby enabling the use of established k-means optimization methods such as the Lloyd's algorithm [Lloyd, 1982]. To apply

Algorithm 1: Topological k-means.

Hyperparameters: K, J, T.
Input: $S = \{D_i\}_{i=1}^n$.
Initialize landscape centroid $M = \{\mu_1, \ldots, \mu_K\}$.
Top-J landscapes are used.
for $t \in 1, \ldots, T$ **do**
  Initialize empty set $S_k = \varnothing$ for all $k \in [K]$.
  for $i \in 1, \ldots, n$ **do**
    $k \leftarrow \pi(\Lambda(D_i) \mid M)$.
    $S_k \leftarrow S_k \cup \{D_i\}$.
  $\mu_k \leftarrow \frac{1}{|S_k|} \sum_{D_i \in S_k} \Lambda(D_i)$ for all $k \in [K]$.
  if *converges* **then**
    break
$c_k^{\text{snap}} \leftarrow \arg\min_{D_i \in S_k} \|\Lambda(D_i) - \mu_k\|_\infty$ for $k \in [K]$.
Output: $C^{\text{snap}} = \{c_1^{\text{snap}}, \ldots, c_K^{\text{snap}}\}$.

the Lloyd's algorithm, we first specify the hyperparameters required for the optimization process: number of clusters K, number of landscapes J, and iteration T. Then, we randomly initialize the landscape centroids $M = \{\mu_1, \ldots, \mu_K\}$. For every iteration $t \in 1, \ldots, T$, we assign each data $D_i \in S$ to the cluster $\pi(\Lambda(D_i)|M)$ and subsequently compute the updated cluster center. The update step computes each cluster center as the pointwise mean over all landscapes assigned to that cluster. Once the algorithm converges, we snap each landscape centroid $\mu_k \in M$ to its closest sample landscape, thereby recovering diagram-space cluster centers via inversion. The procedure is summarized in Algorithm 1.

5 EXPERIMENTS

We next empirically verify the soundness of our method on synthetic data and real-world graph-structured data. Details of the experiment settings are provided in Appendix B.

5.1 SYNTHETIC DATASET

First, we conduct a simple yet illustrative experiment on synthetic data constructed by fixing four ground-truth diagram-space cluster centers, each consisting of three points. Each cluster center is specified as well-separated persistence diagrams that contain three, two, one, and zero prominent off-diagonal points, respectively (see Figure 1). We generate 1000 diagram samples per cluster by perturbing the points in each ground-truth centroid with small Gaussian noise. The landscape centroids are randomly initialized by sampling bounded off-diagonal points to construct arbitrary diagrams, which are then converted to corresponding landscapes. The k-means optimization procedure follows Algorithm 1, and the experiment results are illustrated in Figure 1.

Results. We observe that landscape-based clustering effectively recovers the ground-truth diagram clusters. As shown

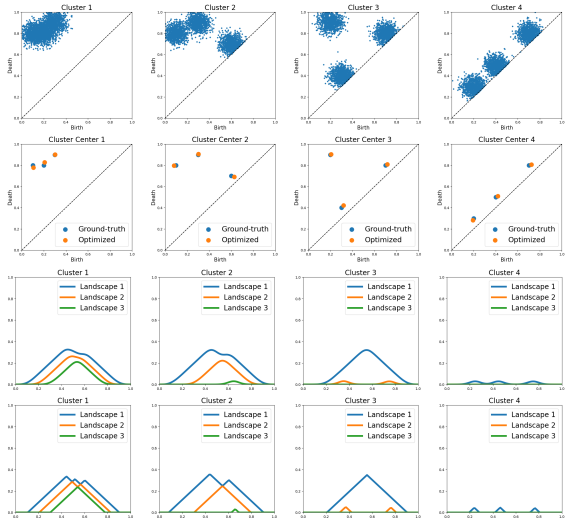

Figure 1: Top row: distribution of 1000 sampled diagrams per cluster. Second row: ground-truth diagram centroid for each cluster (blue) and diagram centroids return by the k-means procedure and snapping (orange). Third row: landscape centroids returned by the k-means procedure. Bottom row: snapped landscape centroids.

in Figure 1, the diagram centroids returned by the Hilbert space k-means procedure remain closely aligned with the ground-truth centroids, even after snapping. A visual comparison of the landscape centroids before and after snapping suggests that the snapping step had only a modest effect, as their overall shapes remain generally consistent. Moreover, by performing clustering in the Hilbert space, we empirically observe rapid convergence, with the algorithm converging in fewer than 10 iterations and incurring minimal computational overhead.

5.2 GEOM-DRUGS DATASET

Across various fields of molecular science, it is often of interest to classify structured data, such as protein bonds or molecules, based on their topological characteristics (e.g., loop counts) [Wee and Jiang, 2025]. To this end, we evaluate our method on the GEOM-Drugs dataset [Axelrod and Gomez-Bombarelli, 2022], which provides graph-structured representations of molecular compounds (see Figure 2). We first sample graphs with one, two, three, and four loops, yielding four distinct clusters with 1000 samples per cluster. We then apply the k-means algorithm as in the previous experiment, using only the 1-dimensional homology features. The experiment results are illustrated in Figure 3.

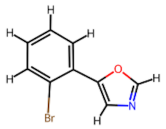

Figure 2: A molecular graph with two loops.

Results. Each optimized diagram center in Figure 3 exhibits

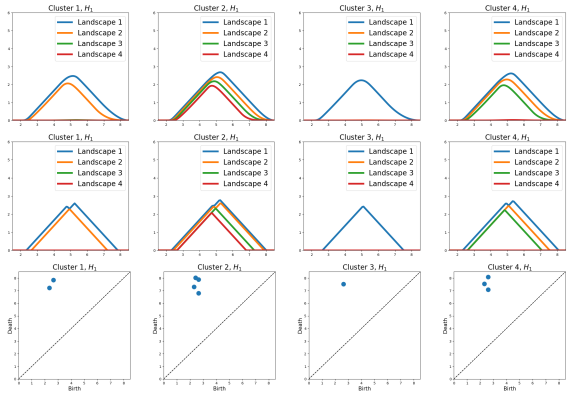

Figure 3: Top row: landscape centroids returned by the k-means procedure. Middle row: snapped landscape centroids. Bottom row: corresponding diagram centroids.

one, two, three, and four 1-dimensional homological features, effectively capturing the salient structure of each cluster. Moreover, the algorithm achieves 98.95% classification accuracy, indicating that the optimized centroids provide reliable representatives of the distinct clusters. As in the previous experiment, we again observe rapid convergence and low runtime.

Overall, the experimental findings validate that clustering persistence landscapes serves as a sound surrogate for directly clustering in the space of persistence diagrams, effectively recovering true diagram centroids under fast convergence rates and low computational costs.

6 CONCLUSION

This work fills a notable gap in the topological clustering literature by providing a principled framework under which persistence landscapes serve as a sound surrogate for k-means clustering in the space of persistence diagrams. We derive explicit conditions under which the landscape embedding is faithful, guaranteeing both label preservation and calibrated control of diagram-space excess risk. Our risk analysis shows that landscape-based objectives govern diagram-space performance with fast rates inherited from Hilbert-space k-means under mild regularity conditions. Algorithmically, the framework legitimizes the use of standard tools such as Lloyd's algorithm, which are not directly applicable in diagram space due to the absence of a canonical averaging operation. We further address the practical issue that landscape centroids produced by k-means are typically not valid landscapes, by introducing a simple data-driven snapping procedure to nearby sampled landscapes and characterizing when the resulting diagram centroids preserve labels and remain near-optimal. Collectively, our framework offers a computationally efficient, theoretically grounded surrogate approach to clustering persistence diagrams.

References

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.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

Pranjal Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. In *APPROX/RANDOM 2012*, volume 7408 of *Lecture Notes in Computer Science*, pages 37–49. Springer, 2012. doi: 10.1007/978-3-642-32512-0_4.

Simon Axelrod and Rafael Gomez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9 (1):185, 2022.

Debanjali Bhattacharya, Neelam Sinha, R Yashwanth, and Amit Chattopadhyay. Image complexity-based fmri-bold visual network categorization across visual datasets using topological descriptors and deep-hybrid learning. *Scientific Reports*, 15(1):36845, 2025.

Gérard Biau, Luc Devroye, and Gábor Lugosi. On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, 54(2):781–790, 2008.

Peter Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16(3):77–102, 2015.

Yueqi Cao, Prudence Leung, and Anthea Monod. k-means clustering for persistent homology. *Advances in Data Analysis and Classification*, 19(1):95–119, 2025.

Mathieu Carrière, Frédéric Chazal, Yuichi Ike, Théo Lacombe, Martin Royer, and Yuhei Umeda. Perslay: A neural network layer for persistence diagrams and new graph topological signatures. In *International Conference on Artificial Intelligence and Statistics*, pages 2786–2796. PMLR, 2020.

Kit C Chan, Umar Islambekov, Alexey Luchinsky, and Rebecca Sanders. A computationally efficient framework for vector representation of persistence diagrams. *Journal of Machine Learning Research*, 23(268):1–33, 2022.

Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence*, 4:108, 2021.

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*, pages 474–483, 2014.

H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. Applied Mathematics. American Mathematical Society, 2010. ISBN 9780821849255.

Jennifer Gamble and Giseon Heo. Exploring uses of persistent homology for statistical analysis of landmark-based shape data. *Journal of Multivariate Analysis*, 101(9): 2184–2199, 2010.

Christoph Hofer, Roland Kwitt, Marc Niethammer, and Andreas Uhl. Deep learning with topological signatures. *Advances in Neural Information Processing Systems*, 30, 2017.

Kwangho Kim, Jisu Kim, and Alessandro Rinaldo. Time series featurization via topological data analysis. *arXiv preprint arXiv:1812.02987*, 2018.

Kwangho Kim, Jisu Kim, Manzil Zaheer, Joon Kim, Frédéric Chazal, and Larry Wasserman. Pllay: Efficient topological layer based on persistent landscapes. *Advances in Neural Information Processing Systems*, 33, 2020.

Genki Kusano, Yasuaki Hiraoka, and Kenji Fukumizu. Persistence weighted gaussian kernel for topological data analysis. In *International conference on machine learning*, pages 2004–2013. PMLR, 2016.

Théo Lacombe, Marco Cuturi, and Steve Oudot. Large scale computation of means and clusters for persistence diagrams using optimal transport. *Advances in Neural Information Processing Systems*, 31, 2018.

Clément Levrard. Fast rates for empirical vector quantization. *Electronic Journal of Statistics*, 7:1716–1746, 2013. doi: 10.1214/13-EJS822.

Clément Levrard. Nonasymptotic bounds for vector quantization in hilbert spaces. *The Annals of Statistics*, 43(2): 592–619, 2015. doi: 10.1214/14-AOS1293.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

Rafail Ostrovsky, Yuval Rabani, Leonard J Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. *Journal of the ACM (JACM)*, 59(6):1–22, 2013.

Jan Reininghaus, Stefan Huber, Ulrich Bauer, and Roland Kwitt. A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4741–4748, 2015.

Bernadette J Stolz, Tegan Emerson, Satu Nahkuri, Mason A Porter, and Heather A Harrington. Topological data analysis of task-based fmri data from experiments on schizophrenia. *Journal of Physics: Complexity*, 2(3): 035006, 2021.

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

JunJie Wee and Jian Jiang. A review of topological data analysis and topological deep learning in molecular sciences. *Journal of Chemical Information and Modeling*, 65(23):12691–12706, 2025.

Appendix

Hajin Lee[1] **Kwangho Kim**[*1]

[1]Department of Statistics, Korea University

A PROOFS

Proof of Theorem 4.1. For notational convenience, we will use the abbreviation $\pi_C(i) \coloneqq \pi(D_i \,|\, C)$, $\pi_\Lambda(i) \coloneqq \pi(\Lambda(D_i) \,|\, \Lambda(C))$, and $\pi_M(i) \coloneqq \pi(\Lambda(D_i) \,|\, M)$.

First, we prove $\pi(\Lambda(D_i) \,|\, \Lambda(C)) = \pi(\Lambda(D_i) \,|\, M)$. By the triangle inequality, we have for all $D_i \in S$ and $k \in [K]$,

$$\|\Lambda(D_i) - \Lambda(c_k)\|_\infty \;\leq\; \|\Lambda(D_i) - \mu_k\|_\infty + \|\Lambda(c_k) - \mu_k\|_\infty,$$

and

$$\|\Lambda(D_i) - \Lambda(c_k)\|_\infty \geq\; \|\Lambda(D_i) - \mu_k\|_\infty - \|\Lambda(c_k) - \mu_k\|_\infty.$$

For cluster $k = \pi_M(i)$, we use the upper bound

$$\begin{aligned}\|\Lambda(D_i) - \Lambda(c_{\pi_M(i)})\|_\infty \;&\leq\; \|\Lambda(D_i) - \mu_{\pi_M(i)}\|_\infty + \|\Lambda(c_{\pi_M(i)}) - \mu_{\pi_M(i)}\|_\infty \\ &=\; \|\Lambda(D_i) - \mu_{\pi_M(i)}\|_\infty + \tau_{\pi_M(i)} \\ &\leq\; \|\Lambda(D_i) - \mu_{\pi_M(i)}\|_\infty + \tau.\end{aligned}$$

since $\|\Lambda(c_k) - \mu_k\|_\infty = \tau_k \leq \tau$ for all k by definition. For clusters $k \neq \pi_M(i)$, we use the lower bound

$$\begin{aligned}\|\Lambda(D_i) - \Lambda(c_k)\|_\infty \;&\geq\; \|\Lambda(D_i) - \mu_k\|_\infty - \|\Lambda(c_k) - \mu_k\|_\infty \\ &=\; \|\Lambda(D_i) - \mu_k\|_\infty - \tau_k \\ &\geq\; \|\Lambda(D_i) - \mu_k\|_\infty - \tau.\end{aligned}$$

Combining the inequalities, we have for any $k \neq \pi_M(i)$,

$$\begin{aligned}\|\Lambda(D_i) - \Lambda(c_k)\|_\infty - \|\Lambda(D_i) - \Lambda(c_{\pi_M(i)})\|_\infty \;&\geq\; \|\Lambda(D_i) - \mu_k\|_\infty - \|\Lambda(D_i) - \mu_{\pi_M(i)}\|_\infty - 2\tau \\ &\geq\; \Delta_i - 2\tau.\end{aligned}$$

If $\Delta_i > 2\tau$ for all i, $\|\Lambda(D_i) - \Lambda(c_k)\|_\infty$ is strictly larger than $\|\Lambda(D_i) - \Lambda(c_{\pi_M(i)})\|_\infty$ for any $k \neq \pi_M(i)$. Thus, we have $\pi(\Lambda(D_i) \,|\, \Lambda(C)) = \pi(\Lambda(D_i) \,|\, M)$ for all $D_i \in S$.

Next, we prove $\pi(D_i \,|\, C) = \pi(\Lambda(D_i) \,|\, \Lambda(C))$ using contradiction. Suppose that for any given $D_i \in S$, it holds that $\pi_C(i) \neq \pi_\Lambda(i)$. Then, by definition, we have

$$\|\Lambda(D_i) - \Lambda(c_{\pi_\Lambda(i)})\|_\infty \leq \|\Lambda(D_i) - \Lambda(c_{\pi_C(i)})\|_\infty.$$

[*]Corresponding author.
[*]Corresponding author.

We can also derive

$$\|\Lambda(D_i) - \Lambda(c_{\pi_C(i)})\|_\infty \le d_B(D_i, c_{\pi_C(i)})$$

from (1), and

$$\alpha_n d_B(D_i, c_{\pi_\Lambda(i)}) \le \|\Lambda(D_i) - \Lambda(c_{\pi_\Lambda(i)})\|_\infty$$

by definition. Combining the above inequalities yields

$$\alpha_n d_B(D_i, c_{\pi_\Lambda(i)}) \le d_B(D_i, c_{\pi_C(i)}) \quad \Rightarrow \quad \frac{d_B(D_i, c_{\pi_\Lambda(i)})}{d_B(D_i, c_{\pi_C(i)})} \le \frac{1}{\alpha_n},$$

and by definition,

$$r_i \le \frac{d_B(D_i, c_{\pi_\Lambda(i)})}{d_B(D_i, c_{\pi_C(i)})}.$$

However, if $r_i > 1/\alpha_n$, the two inequalities cannot be satisfied simultaneously. Thus, by contradiction, we have $\pi(D_i \,|\, C) = \pi(\Lambda(D_i) \,|\, \Lambda(C))$.

Combining the two intermediate results completes the proof:

$$\pi(D_i \,|\, C) = \pi(\Lambda(D_i) \,|\, \Lambda(C)) = \pi(\Lambda(D_i) \,|\, M).$$

□

Proof of Theorem 4.2. For notational convenience, we again use the abbreviation $\pi_C(i) \coloneqq \pi(D_i \,|\, C)$, $\pi_\Lambda(i) \coloneqq \pi(\Lambda(D_i) \,|\, \Lambda(C))$, and $\pi_M(i) \coloneqq \pi(\Lambda(D_i) \,|\, M)$.

We first prove $\pi(\Lambda(D_i) \,|\, \Lambda(C)) = \pi(\Lambda(D_i) \,|\, M)$. From the proof of Theorem 4.1, we have

$$\|\Lambda(D_i) - \Lambda(c_{\pi_M(i)})\|_\infty \;\le\; \|\Lambda(D_i) - \mu_{\pi_M(i)}\|_\infty + \tau_{\pi_M(i)}$$

for cluster $k = \pi_M(i)$, and

$$\|\Lambda(D_i) - \Lambda(c_k)\|_\infty \;\ge\; \|\Lambda(D_i) - \mu_k\|_\infty - \tau_k$$

for clusters $k \neq \pi_M(i)$. Combining the inequalities, we have for any $k \neq \pi_M(i)$,

$$\begin{aligned}\|\Lambda(D_i) - \Lambda(c_k)\|_\infty - \|\Lambda(D_i) - \Lambda(c_{\pi_M(i)})\|_\infty \;&\ge\; \|\Lambda(D_i) - \mu_k\|_\infty - \|\Lambda(D_i) - \mu_{\pi_M(i)}\|_\infty - \tau_{\pi_M(i)} - \tau_k \\ &=\; \Delta_{ik} - \tau_{\pi_M(i)} - \tau_k.\end{aligned}$$

If $\Delta_{ik} > \tau_{\pi_M(i)} + \tau_k$ for all i and $k \neq \pi_M(i)$, $\|\Lambda(D_i) - \Lambda(c_k)\|_\infty$ is strictly larger than $\|\Lambda(D_i) - \Lambda(c_{\pi_M(i)})\|_\infty$ for any $k \neq \pi_M(i)$. Thus, we have $\pi(\Lambda(D_i) \,|\, \Lambda(C)) = \pi(\Lambda(D_i) \,|\, M)$ for all $D_i \in S$.

Next, we prove $\pi(D_i \,|\, C) = \pi(\Lambda(D_i) \,|\, \Lambda(C))$. By definition, $\|\Lambda(D_i) - \Lambda(c_k)\|_\infty = \alpha_{ik}\, d_B(D_i, c_k)$ for $k \neq \pi_C(i)$, while (1) gives $\|\Lambda(D_i) - \Lambda(c_{\pi_C(i)})\|_\infty \le d_B(D_i, c_{\pi_C(i)})$. The assumption

$$d_B(D_i, c_{\pi_C(i)}) \;<\; \alpha_{ik}\, d_B(D_i, c_k), \quad \text{for every } D_i \in S \text{ and } k \neq \pi_C(i),$$

ensures that the distance from $\Lambda(D_i)$ to $\Lambda(c_{\pi_C(i)})$ remains strictly smaller in the landscape space than to any competing center. □

Proof of Corollary 4.3. The result follows immediately from Theorem 4.1 and 4.2 □

Proof of Theorem 4.4. From (1), for any diagram D and any center $c_k \in C$,

$$\|\Lambda(D) - \Lambda(c_k)\|_\infty \;\le\; d_B(D, c_k).$$

Taking the minimum over $k \in [K]$ and subsequently applying the nondecreasing map ρ does not change the inequality:

$$\rho\Big(\min_{k\in[K]} \|\Lambda(D) - \Lambda(c_k)\|_\infty\Big) \le \rho\Big(\min_{k\in[K]} d_B(D, c_k)\Big).$$

Averaging this pointwise inequality over $D = D_i$ yields $\widehat{R}_\Lambda(\Lambda(C)) \le \widehat{R}_{\mathcal{D}}(C)$.

For the population risk, an identical claim holds for the expectation in place of the empirical average. Thus, $R_\Lambda(\Lambda(C)) \le R_{\mathcal{D}}(C)$. □

Proof of Theorem 4.5. For any given $C \in \mathcal{C}_K$, we have

$$\alpha d_B(D, c_{\pi(\Lambda(D)|\Lambda(C))}) \;\le\; \|\Lambda(D) - \Lambda(c_{\pi(\Lambda(D)|\Lambda(C))})\|_\infty$$

by Assumption 2. It also holds that

$$\alpha\Big(\min_{k\in[K]} d_B(D, c_k)\Big) \;\le\; \alpha d_B(D, c_{\pi(\Lambda(D)|\Lambda(C))}) \;\le\; \|\Lambda(D) - \Lambda(c_{\pi(\Lambda(D)|\Lambda(C))})\|_\infty = \min_{k\in[K]} \|\Lambda(D) - \Lambda(c_k)\|_\infty,$$

and applying the nondecreasing map $\rho : t \mapsto t^q$ does not change the inequality:

$$\alpha^q\Big(\min_{k\in[K]} d_B(D, c_k)\Big)^q \le \Big(\min_{k\in[K]} \|\Lambda(D) - \Lambda(c_k)\|_\infty\Big)^q.$$

Finally, taking the expectation with respect to D and rearranging yields the first desired result:

$$\alpha^q R_{\mathcal{D}}(C) \le R_\Lambda(\Lambda(C)) \quad \Rightarrow \quad R_{\mathcal{D}}(C) \le \alpha^{-q} R_\Lambda(\Lambda(C)).$$

Next, for the excess risk bound, first observe that

$$R_\Lambda(\Lambda(C^*_\Lambda)) \le R_\Lambda(\Lambda(C^*)) \le R_{\mathcal{D}}(C^*),$$

since C^*_Λ minimizes $R_\Lambda(\Lambda(C))$, and $R_\Lambda(\Lambda(C)) \le R_{\mathcal{D}}(C)$ pointwise by Theorem 4.4. Thus, we have

$$R_{\mathcal{D}}(C) - R_{\mathcal{D}}(C^*) \le R_{\mathcal{D}}(C) - R_\Lambda(\Lambda(C^*_\Lambda)) \le \alpha^{-q} R_\Lambda(\Lambda(C)) - R_\Lambda(\Lambda(C^*_\Lambda)).$$

Adding and subtracting $\alpha^{-q} R_\Lambda(\Lambda(C^*_\Lambda))$ to the upper bound gives the desired result:

$$\begin{aligned} R_{\mathcal{D}}(C) - R_{\mathcal{D}}(C^*) &\le \alpha^{-q} R_\Lambda(\Lambda(C)) - \alpha^{-q} R_\Lambda(\Lambda(C^*_\Lambda)) + \alpha^{-q} R_\Lambda(\Lambda(C^*_\Lambda))) - R_\Lambda(\Lambda(C^*_\Lambda)) \\ &= \alpha^{-q}\{R_\Lambda(\Lambda(C)) - R_\Lambda(\Lambda(C^*_\Lambda))\} + (\alpha^{-q} - 1) R_\Lambda(\Lambda(C^*_\Lambda)). \end{aligned}$$

□

Proof of Theorem 4.6. We first prove the diagram-space excess risk upper bound for $\widehat{C}^n_\Lambda$ in (6). For any $C \in \mathcal{C}_K$, we have

$$R_\Lambda(\Lambda(C)) \;\le\; \widehat{R}_\Lambda(\Lambda(C)) + \Gamma_n, \qquad R_\Lambda(\Lambda(C)) \;\ge\; \widehat{R}_\Lambda(\Lambda(C)) - \Gamma_n,$$

with probability at least $1 - \delta_n$ from (5). In particular, for $C = \widehat{C}^n_\Lambda$ and $C = C^*_\Lambda$,

$$R_\Lambda(\Lambda(\widehat{C}^n_\Lambda)) \;\le\; \widehat{R}_\Lambda(\Lambda(\widehat{C}^n_\Lambda)) + \Gamma_n, \qquad R_\Lambda(\Lambda(C^*_\Lambda)) \;\ge\; \widehat{R}_\Lambda(\Lambda(C^*_\Lambda)) - \Gamma_n.$$

By the approximate empirical optimality of $\widehat{C}^n_\Lambda$,

$$\widehat{R}_\Lambda(\Lambda(\widehat{C}^n_\Lambda)) \;\le\; \widehat{R}_\Lambda(\Lambda(C^*_\Lambda)) + \varepsilon_n.$$

Putting together these inequalities yields that, with probability at least $1 - \delta_n$,

$$R_\Lambda(\Lambda(\widehat{C}^n_\Lambda)) - R_\Lambda(\Lambda(C^*_\Lambda)) \;\le\; \big\{\widehat{R}_\Lambda(\Lambda(C^*_\Lambda)) + \varepsilon_n + \Gamma_n\big\} - \big\{\widehat{R}_\Lambda(\Lambda(C^*_\Lambda)) - \Gamma_n\big\} = \varepsilon_n + 2\Gamma_n.$$

Combined with Theorem 4.5, we have

$$\begin{aligned} R_{\mathcal{D}}(\widehat{C}^n_\Lambda) - R_{\mathcal{D}}(C^*) &\le \alpha^{-q}\big(R_\Lambda(\Lambda(\widehat{C}^n_\Lambda)) - R_\Lambda(\Lambda(C^*_\Lambda))\big) + (\alpha^{-q} - 1)\, R_\Lambda(\Lambda(C^*_\Lambda)) \\ &\le \alpha^{-q}\big(\varepsilon_n + 2\Gamma_n\big) + (\alpha^{-q} - 1)\, R_\Lambda(\Lambda(C^*_\Lambda)). \end{aligned}$$

Next, we prove that $R_{\mathcal{D}}(\widehat{C}^n_\Lambda)$ converges in probability to $R_{\mathcal{D}}(C^*)$ with irreducible structural bias $(\alpha^{-q} - 1)\, R_\Lambda(\Lambda(C^*_\Lambda))$. If $\varepsilon_n \to 0$, $\delta_n \to 0$ and $\Gamma_n \to 0$, the stochastic term $\alpha^{-q}(\varepsilon_n + 2\Gamma_n)$ converges to 0 with probability $1 - \delta_n \to 1$. Thus, $R_{\mathcal{D}}(\widehat{C}^n_\Lambda)$ converges to $R_{\mathcal{D}}(C^*) + (\alpha^{-q} - 1)\, R_\Lambda(\Lambda(C^*_\Lambda))$ in probability. □

Proof of Corollary 4.7. Recall from the proof of Theorem 4.6 that

$$R_{\mathcal{D}}(\widehat{C}^n_\Lambda) - R_{\mathcal{D}}(C^*) \;\le\; \alpha^{-q}\big(R_\Lambda(\Lambda(\widehat{C}^n_\Lambda)) - R_\Lambda(\Lambda(C^*_\Lambda))\big) + (\alpha^{-q}-1)\, R_\Lambda(\Lambda(C^*_\Lambda)).$$

Combining this with the assumed rate $R_\Lambda(\Lambda(\widehat{C}^n_\Lambda)) - R_\Lambda(\Lambda(C^*_\Lambda)) = O_p(\varphi_n)$, we immediately get

$$R_{\mathcal{D}}(\widehat{C}^n_\Lambda) - R_{\mathcal{D}}(C^*) \;=\; O_p\big(\alpha^{-q}\varphi_n\big) + (\alpha^{-q}-1)\, R_\Lambda(\Lambda(C^*_\Lambda)),$$

since $(\alpha^{-q}-1)\, R_\Lambda(\Lambda(C^*_\Lambda))$ is deterministic and does not affect the stochastic order. □

Proof of Corollary 4.8. Using $\rho(t) = t^q$ with $q = 2$, the assumption

$$R_\Lambda(\Lambda(\widehat{C}^n_\Lambda)) - R_\Lambda(\Lambda(C^*_\Lambda)) = O_p(\varphi_n)$$

directly implies the first result

$$R_{\mathcal{D}}(\widehat{C}^n_\Lambda) - R_{\mathcal{D}}(C^*) = O_p\big(\alpha^{-2}\varphi_n\big) + (\alpha^{-2}-1)\, R_\Lambda(\Lambda(C^*_\Lambda)),$$

from Corollary 4.7.

Otherwise, if $\widehat{C}^n_\Lambda$ is ε_n-optimal empirically and the uniform deviation Γ_n is of order $O_p(\varphi_n)$, we have

$$R_\Lambda(\Lambda(\widehat{C}^n_\Lambda)) - R_\Lambda(\Lambda(C^*_\Lambda)) \;=\; O_p(\varepsilon_n) + O_p(\varphi_n),$$

since

$$R_\Lambda(\Lambda(\widehat{C}^n_\Lambda)) - R_\Lambda(\Lambda(C^*_\Lambda)) \;\le\; \varepsilon_n + 2\Gamma_n,$$

from the proof of Theorem 4.6. Thus, the same bound with ε_n explicitly appearing carries through to the diagram-space side, yielding the stated order $O_p(\alpha^{-2}\varepsilon_n) + O_p(\alpha^{-2}\varphi_n)$ plus the problem-dependent bias term $(\alpha^{-2}-1)\, R_\Lambda(\Lambda(C^*_\Lambda))$. □

Proof of Lemma 4.9. Let $\pi_\Lambda(i) \coloneqq \pi(\Lambda(D_i) \,|\, \Lambda(C^{\mathrm{snap}}))$, and $\pi_M(i) \coloneqq \pi(\Lambda(D_i) \,|\, M)$. By triangle inequality and the snapping bound $\|\Lambda(c_k) - \mu_k\|_\infty \le \tau_k \le \tau$, we have for all $k \in [K]$,

$$\big|\|\Lambda(D_i) - \Lambda(c_k)\|_\infty - \|\Lambda(D_i) - \mu_k\|_\infty\big| \;\le\; \|\Lambda(c_k) - \mu_k\|_\infty \;\le\; \tau.$$

Thus, we have

$$\|\Lambda(D_i) - \Lambda(c_k)\|_\infty \le\; \|\Lambda(D_i) - \mu_k\|_\infty + \tau \quad \Rightarrow \|\Lambda(D_i) - \Lambda(c_{\pi_\Lambda(i)})\|_\infty \le\; \|\Lambda(D_i) - \mu_{\pi_M(i)}\|_\infty + \tau,$$

and

$$\|\Lambda(D_i) - \mu_k\|_\infty \le \|\Lambda(D_i) - \Lambda(c_k)\|_\infty + \tau \quad \Rightarrow \|\Lambda(D_i) - \mu_{\pi_M(i)}\|_\infty \le \|\Lambda(D_i) - \Lambda(c_{\pi_\Lambda(i)})\|_\infty + \tau,$$

by first taking the minimum over k on the left hand side and subsequently taking the minimum over k on the right hand side. Combining the above inequalities yields

$$\big|\|\Lambda(D_i) - \Lambda(c_{\pi_\Lambda(i)})\|_\infty - \|\Lambda(D_i) - \mu_{\pi_M(i)}\|_\infty\big| \;\le\; \tau.$$

By the definition of B, we have $\|\Lambda(D_i) - \Lambda(c_{\pi_\Lambda(i)})\|_\infty,\ \|\Lambda(D_i) - \mu_{\pi_M(i)}\|_\infty \in [0, B]$, and $\rho(t) = t^q$ is L_q-Lipschitz on $[0, B]$, so

$$\begin{aligned}
&\big|\big\{\|\Lambda(D_i) - \Lambda(c_{\pi_\Lambda(i)})\|_\infty\big\}^q - \big\{\|\Lambda(D_i) - \mu_{\pi_M(i)}\|_\infty\big\}^q\big| \\
&\le\; L_q\big|\|\Lambda(D_i) - \Lambda(c_{\pi_\Lambda(i)})\|_\infty - \|\Lambda(D_i) - \mu_{\pi_M(i)}\|_\infty\big| \\
&\le\; L_q\tau.
\end{aligned}$$

Given that $|\frac{1}{n}\sum_{i=1} x_i| \;\le\; \frac{1}{n}\sum_{i=1}|x_i|$, the above inequality further yields

$$\big|\widehat{R}_\Lambda(\Lambda(C^{\mathrm{snap}})) - \widehat{R}_\Lambda(\widehat{M}^n)\big| \;\le\; L_q\tau.$$

Next, note that for any $C' = \{c'_1, \ldots, c'_K\}$, we have $\widehat{R}_\Lambda(\Lambda(C')) = \widehat{R}_\Lambda(M')$ by setting $M' = \{\Lambda(c'_1), \ldots, \Lambda(c'_K)\}$. Thus,

$$\inf_{M' \in \mathcal{M}_K} \widehat{R}_\Lambda(M') \;\leq\; \inf_{C' \in \mathcal{C}_K} \widehat{R}_\Lambda(\Lambda(C')).$$

Combining the two inequalities above with the assumption $\widehat{R}_\Lambda(\widehat{M}^n) \;\leq\; \inf_{M \in \mathcal{M}_K} \widehat{R}_\Lambda(M) \;+\; \gamma_n$, we have

$$\widehat{R}_\Lambda(\Lambda(C^{\mathrm{snap}})) \;\leq\; \widehat{R}_\Lambda(\widehat{M}^n) + L_q\tau \leq \inf_{C \in \mathcal{C}_K} \widehat{R}_\Lambda(\Lambda(C)) + L_q\tau + \gamma_n.$$

By setting $\varepsilon_n := L_q\,\tau + \gamma_n$, we get

$$\widehat{R}_\Lambda(\Lambda(C^{\mathrm{snap}})) \;\leq\; \inf_{C \in \mathcal{C}_K} \widehat{R}_\Lambda(\Lambda(C)) + \varepsilon_n.$$

□

B EXPERIMENT DETAILS

B.1 SYNTHETIC DATA

The four ground-truth diagram-space cluster centers consisting of three persistence points are set as

1. ((0.1, 0.8), (0.2, 0.8), (0.3, 0.9)),
2. ((0.1, 0.8), (0.3, 0.9), (0.6, 0.7)),
3. ((0.2, 0.9), (0.3, 0.4), (0.7, 0.8)),
4. ((0.2, 0.3), (0.4, 0.5), (0.7, 0.8)),

respectively. 1000 diagrams are sampled per cluster by perturbing the points in each ground-truth centroid with small Gaussian noise (mean 0, variance 0.05), subject to the constraint that all points remain above the diagonal and are bounded within $[0, 1]^2$. We set the number of clusters to $K = 4$ and use $J = 3$ landscapes, computed on the interval $[0, 1]$ with 100 discretized points. Landscape centroids are randomly initialized by first sampling birth-death pairs uniformly from $(0.01, 0.99)$ such that the birth time is smaller than the death time, and subsequently mapping them to corresponding landscapes. The k-means optimization procedure converges in seven iterations with a runtime of 0.2 seconds.

B.2 GEOM-DRUGS DATA

We first sample graphs with one, two, three, and four loops, yielding four distinct clusters with 1000 samples per cluster. Each graph node has six features, and we assign a scalar node weight as a randomly sampled convex combination of these feature values. Each edge is assigned a weight given by the maximum of its adjacent node weights. Since 1-dimensional homology features have infinite death times, we replace them with values sampled uniformly from $[6.5, , 8.5]$. We set the number of clusters to $K = 4$ and use $J = 4$ landscapes, computed on the interval $[1.5, 8.5]$ with 100 discretized points. Landscape centroids are initialized as in the previous experiment, and the k-means optimization procedure converges in seven iterations with a runtime of 0.2 seconds.