ON SUBSAMPLING AND INFERENCE FOR MULTIPARAMETER PERSISTENCE HOMOLOGY

Vinoth Nandakumar *
Trustworthy Machine Learning Lab, Department of Computer Science
University of Sydney, Australia
vinoth.nandakumar@sydney.edu.au

ABSTRACT

In topological data analysis, multi-parameter persistence homology is a framework for extracting topological information for point cloud datasets equipped with multiple filtrations. However, existing algorithms become computationally expensive for large datasets, limiting their applicability. In the single-parameter case, subsampling algorithms have been used to reduce the time complexity of computing persistence homology. Convergence properties of the persistence barcodes have also been established in this setting. We extend these results to the multiparameter persistence homology, and develop subsampling algorithms can be used to approximate the fibered barcode in this setting. We conduct experiments on the point cloud dataset ModelNet to demonstrate the efficiency of these algorithms.

1 INTRODUCTION

Topological data analysis is an rapidly growing field that producing statistical summaries of datasets by leveraging their underlying topological structure, inspired by the theory of homology from algebraic topology. Notable applications include identifying biological datasets [21] and image segmentation [19].

One-parameter persistent homology ([13], [25]) extracts topological information from a dataset by constructing a family of simplicial complexes, and studying how the homology varies. The resulting statistical summary is presented as a persistence barcode. In many real-world applications, there are multiple parameters attached to the dataset and it is important to incorporate this additional information. However, the entire combinatorial structure of the multi-parameter persistence module cannot be encoded in a simple diagram analogous to the persistence barcode ([4]). The fibered barcode ([5]) is a statistical summary that can be associated to a dataset equipped with a multi-parameter filtration. While it is not an invariant, it does captures much of the interesting topological information.

These algorithms are computationally expensive for large datasets, and it is desirable to have efficient methods of approximating the statistical summaries. An approach proposed by [8] in the one-parameter setting is to choose several subsamples from the dataset, compute their persistence landscapes and combine the information to produce an approximation. These subsampling algorithms build on earlier work [9] establishing convergence properties of the persistence homology for samples drawn from a measure on a compact metric space, and use the language of persistence landscapes (following [3]). The key technical step required is the stability of the persistence barcode with respect to the Haussdorff distance.

Our contributions. In the present work, we study subsampling algorithms for multiparameter persistence homology and analyze their performance. We focus on two common multiparameter filtrations - the multi-level set filtration, and the degree-RIPS filtration. First we study the convergence properties of the fibered barcode for samples drawn from a measure μ on a compact metric space M, extending the main results of [9] to both of these filtrations.

Outline. Section 2 gives an overview of the background needed on multiparameter persistence homology, and the stability results. Section 3 establishes convergence results for the fibered barcode

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies. Funding acknowledgements go at the end of the paper.

of finite samples drawn from a measure on a compact metric space, with respect to the matching distance. Section 4 presents the subsampling algorithms, and analyzes their theoretical performance. Section 5 provides experimental results on a synthetic dataset, and on the point cloud dataset ModelNet. All proofs are given in the Appendix.

2 BACKGROUND MATERIAL

This section introduces the notations that will be used throughout the paper, and gives an expository overview of the results about persistence homology that will be needed. None of the results in this section are new.

Let $\mathbf{M}$ be a finite metric space. In this paper we focus on two-parameter persistence homology using the language of bi-filtered simplicial complexes. A bi-filtered simplicial complex built on $\mathbf{M}$ is a family of simplicial complexes $\{K_{s,t}\}_{s,t\geq 0}$, equipped with inclusions $K_{s,t} \to K_{s',t'}$ if $s \leq s'$ and $t \leq t'$ (these inclusions satisfy the natural compatibilities). First we recall the definition of the fibered barcode for multi-parameter filtrations, following Section 1.5 of [20] (see Appendix A for an exposition of persistence barcodes in the single-parameter setting).

Definition 2.1. *Let $\mathcal{L}$ be the space of all affine lines in $\mathbb{R}^2$ with non-negative slope. Given $L \in \mathcal{L}$ with equation $y = l_1 x + l_2$, define*

$$K_{s,L} := K_{s,l_1 s + l_2}$$

Denote by $\beta_L(K)$ the persistence barcode of the filtered simplicial complex $\{K_{s,L}\}_{s\geq 0}$. The fibered barcode $\beta(K)$ is the function with domain $\mathcal{L}$ that sends L to $\beta_L(K)$.

Now recall the following of the matching distance between two fibered barcodes, following [5] (see also Section 3 of [16]):

Definition 2.2. *Given two fibered barcodes β_1 and β_2, the matching distance $d_M(\beta_1, \beta_2)$ between them is defined as follows (here $L \in \mathcal{L}$ as above, and $w(l)$ is its "weight"):*

$$w(L) = \frac{1}{\sqrt{1 + max(l_1, \frac{1}{l_1})^2}}$$

$$d_M(\beta_1, \beta_2) := sup_{L\in\mathcal{L}} w(L) d_b(\beta_{1,L}, \beta_{2,L})$$

We now define a multi-level set filtration as follows. Our input datum consists of a finite metric space $\mathbf{M}$ with distance function $d_{\mathbf{M}} : \mathbf{M} \times \mathbf{M} \to \mathbb{R}$, and a continuous functional $f : \mathbf{M} \to \mathbb{R}$.

Definition 2.3. *We define a two-parameter filtration as follows. Given a finite subset $U \subset \boldsymbol{M}$, let the Vietoris-Rips complex $\boldsymbol{Rips}^f_{\alpha,\beta}(U)$ consist of all simplicies $[x_1, \cdots, x_k]$ such that $d_M(x_i, x_j) \leq \alpha$ and $f(x_i) \leq \beta$ for all $1 \leq i, j \leq k$. Let $\beta_f(M)$ denote its fibered barcode.*

Given two subsets M_1, M_2 of $\mathbf{M}$, recall that the Haussdorff distance can be defined as follows. These definitions can be extended to compact metric spaces $\mathbf{M}$; see Appendix A for more details.

$$d_H(M_1, M_2) = \max\{ \sup_{m_1\in M_1} \inf_{m_2\in M_2} d(m_1, m_2), \sup_{m_2\in M_2} \inf_{m_1\in M_1} d(m_1, m_2)\}$$

3 SUBSAMPLING ALGORITHMS FOR TWO-PARAMETER PERSISTENCE HOMOLOGY

In this section we present algorithms for approximating the fibered barcode by subsampling. We are primarily interested in the following scenario: suppose we have a dataset $\mathbb{X}$ consisting of N points, and we seek to compute its fibered barcode. Computing it precisely is often infeasible when N is large, so instead we proceed by drawing smaller subsamples. We present two algorithms for subsampling, extending the techniques from [8] in the single-parameter setting.

Keeping the notation from the previous section, we now describe two algorithms for approximating the fibered barcode $\beta_f(\mathbb{X})$ obtained from a finite dataset $\mathbb{X}$, and a function $f : \mathbb{X} \to \mathbb{R}$. Algorithm

Algorithm 1 Closest subsample algorithm for multi-level set filtrations

Input: dataset $\mathbb{X}$, function $f : \mathbb{X} \to \mathbb{R}, m, n$
for $i = 0$ **to** $m - 1$ **do**
  Let $\mathbb{X}_{[i]}$ be a randomly chosen sample of $\mathbb{X}$ with size n; let $d_i = d_H(\mathbb{X}_{[i],f}, \mathbb{X}_f)$
end for
$j = \underset{1 \leq i \leq m}{\arg\min}\, d_H(\mathbb{X}, \mathbb{X}_i); \mathbb{X}' = \mathbb{X}_j;\ \epsilon = d_H(\mathbb{X}, \mathbb{X}')$
Output: subset $\mathbb{X}'$, such that $|\mathbb{X}'| = m$ and $d_M(\beta_f(\mathbb{X}), \beta_f(\mathbb{X}')) < 2\epsilon$

1 is an extension of the closest subsample approach from [8]. The Haussdorff distance between finite metric spaces can be computed efficiently (for instance, using the SciPy "directed Haussdorff" package). Experimental results from Section 4.1 indicate that for some real-world datasets, Algorithm 2 is more effective.

Algorithm 2 Haussdorff subsampling algorithm

Input: finite metric space $\mathbb{M}$, error threshold $\epsilon > 0$
Let $\mathbb{M}' = \emptyset, V = \mathbb{X}$
while $V \neq \emptyset$ **do**
  Pick $x \in V$ randomly.
  Let $\mathbb{M}' = \mathbb{M}' \cup x$
  for $y \in V$: **do**
    if $d(y, x) < \epsilon$: **then**
      $V \leftarrow V - \{y\}$
    end if
  end for
end while
Output: subset $\mathbb{X}'$, such that $|\mathbb{X}'| = m$ and $d_M(\beta_f(\mathbb{X}), \beta_f(\mathbb{X}')) < 2\epsilon$

4 EXPERIMENTS

In this section, we present experiments that illustrate the practicality of the subsampling algorithm from the previous section on both synthetic and real datasets. In both cases, the subsampling algorithm in the previous section can be used to quickly approximate the fibered barcode, within a reasonable margin of error. All experiments in this section were performed using the package Gudhi on Google CoLab Pro with a V100 GPU and 25GB RAM.

4.1 3D POINT CLOUD DATASETS

ModelNet10 is a 3D point cloud classification datasets consists of 3991 training samples and 908 test instances belonging to 10 classes of objects (see [24]). We select one poses from three different classes: "Toilet", "Chair" and "Bed". For $0 \leq i \leq 2$, let X_i consist of a $N = 1000$ points chosen from the training samples. The following example illustrates the efficacy of Algorithm 2 in this context (specifically in approximating the longest intervals of the persistent barcodes of $\beta_f(X_i)$ restricted to the line $y = x$). It is unclear if the topological information obtained in this manner can discriminate between classes in ModelNet10.

Example 4.1. *First we rescale the data points so that they are centered at zero, and each of the three components has unit variance. Let $f : X_i \to \mathbb{R}$ be the function defined by $f(x) = \frac{||x||}{10}$ for each point $x \in X_i$. Consider the bi-filtered simplicial complex on X_i arising from the multi-level set construction, and let $\beta_f(X_i)$ denote its fibered barcode. Our objective in this example is to use the algorithms in Section 3 to approximate $\beta_f(X_i)$, queried along a line L: $y = x$. Computing these persistence barcode $\beta_{f,L}(X_i)$ exactly requires more than 310 seconds for each of the three point clouds; see Figure 1 for diagrams of the barcodes obtained. When using Algorithm 1, with $m = 100$ subsamples of $n = 400$ points, we found that the approximate barcodes obtained are too coarse to retain much useful information. When we use the Haussdorff subsampling algorithm with $\epsilon = 0.2$, we found that the longest intervals of the barcodes $\beta_{f,L}(X_i')$ closely resemble those of $\beta_{f,L}(X_i)$*

Table 1: Results of the subsampling algorithm for ModelNet. The "Class" column specifies which of the three point clouds are being used. The "Intervals" column lists all intervals in the persistent barcode $\beta_{f,L}(X_i)$ of length at least 0.4. The "Intervals (Sample)" column lists all intervals in the persistent barcode $\beta_{f,L}(X_i')$ of length at least 0.4, where X_i' is the subsample computed using Algorithm 2. The "Time" column specifies the number of seconds needed for both of these computations (the number in brackets refers to the number of points in the subsample).

CLASS	INTERVALS	INTERVALS (SAMPLE)	TIME
TOILET	[0.349, 0.750]	[0.361, 0.810]	27S (486 PTS)
			311S (1000 PTS)
CHAIR	[0.681 1.21]	[0.714 1.26]	5S (284 PTS)
	[0.339 1.71]	[0.486 1.76]	312S (1000 PTS)
BED	[0.353 0.768]	[0.436 0.845]	19S (434 PTS)
	[0.350 1.076]	[0.4 1.08]	312S (1000 PTS)

(specifically, the intervals whose length is at least 0.4). The results are presented in Table 1. Note however that most of the shorter intervals in $\beta_{f,L}(X_i)$ cannot be reconstructed by subsampling in this fashion.

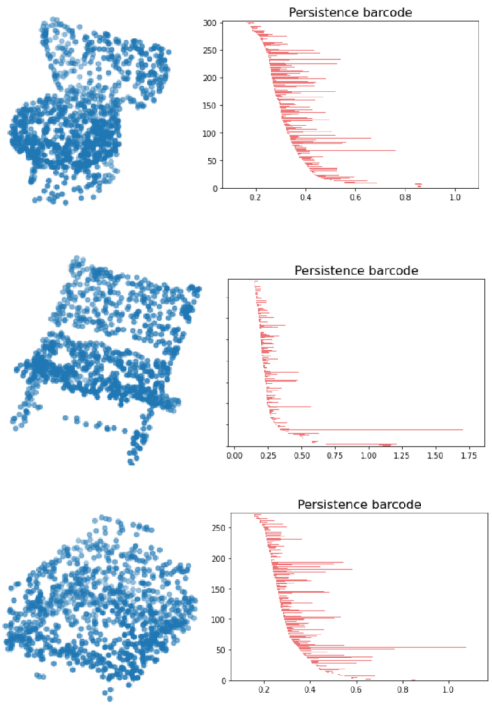

Figure 1: Point clouds from ModelNet.

5 DISCUSSION AND CONCLUSION

In this work, we developed statistical techniques for analyzing multiparameter persistence homology. In particular, we establish convergence results the fibered barcode for a dataset (equipped with a multi-level set filtration) and analyze subsampling algorithms for efficiently approximating this quantity. This extends earlier work of [9] and [8] in the one-parameter setting. In future work,

we will investigate improvements to the subsampling algorithms using multiparameter persistence landscapes.

REFERENCES

[1] Blelloch, Guy; Fineman, Jeremy; Shun, Julian, Greedy sequential maximal independent set and matching are parallel on average, Proceedings of the twenty-fourth annual ACM symposium on Parallelism in algorithms and architectures. ACM, 308–317, 2012

[2] A. Blumberg, M. Lesnick, Stability of 2-Parameter Persistent Homology. arXiv pre-print 2010:09628, 2020

[3] Bubenik, P. Statistical topological data analysis using persistence landscapes. Journal of Machine Learning Research, 16 (2015), 77-102.

[4] Carlsson, G., Zomorodian, A. The theory of multidimensional persistence. Discrete & Computational Geometry, 42(1):71–93, 2009.

[5] Cerri A., Fabio, B., Ferri, M., Frosini, P., Landi, C. Betti numbers in multidimensional persistent homology are stable functions. Mathematical Methods in the Applied Sciences, 36(12):1543–1557, 2013.

[6] Chazal, F., Fasy, B., Lecci, F., Rinaldo, A. and Wasserman, L. Stochastic Convergence of Persistence Landscapes and Silhouettes. Proceedings of the thirtieth annual symposium on Computational geometry, June 2014, pp.474–483

[7] Chazal, F., Silva, V., Oudot, S. Persistence stability for geometric complexes, Geometriae Dedicata, Springer Verlag, 2014, 173, pp.193-214.

[8] F. Chazal, B.Fasy, F.Lecci, B. Michel., A.Rinaldo, L.Wasserman; Subsampling methods for persistent homology Proceedings of the 32nd International Conference on Machine Learning, Vol. 37, 2015.

[9] F. Chazal, M. Glisse, C. Labruere, B. Michel. Convergence rates for persistence diagram estimation in topological data analysis Journal of Machine Learning Research (JMLR), Vol. 16, p. 3603-3635, Dec. 2015.

[10] Cuevas A. and Rodriguez-Casal, A., On boundary estimation, Advances in Applied Probability, 36(2):340-354, 2004.

[11] Cohen-Steiner, D., Edelsbrunner, H. and Harer, J. Stability of persistence diagrams. Discrete & Computational Geometry, 37(1):103-120, 2007.

[12] Dudley R., The speed of mean Glivenko-Cantelli convergence, Ann. Math. Statist. 40 (1969), no. 1, 40-50

[13] Edelsbrunner, H., Letscher, D., Zomorodian, A. Topological persistence and simpli cation. Discrete & Computational Geometry, 28:511-533, 2002.

[14] Fasy, B., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. Confidence sets for persistence diagrams. The Annals of Statistics, 42(6):2301–2339, 2014b.

[15] Kalisnik, S., Lehn, C. and Limic, V. Geometric and probabilistic limit theorems in topological data analysis arXiv pre-print 1903.00470, 2019.

[16] Kerber, M., Lesnick, M. and Oudot, S. Exact Computation of the Matching Distance on 2-Parameter Persistence Modules. 35th International Symposium on Computational Geometry, SoCG (June 18-21, 2019), Oregon, USA

[17] Kerber, M. and Rolle, A., Fast Minimal Presentations of Bi-graded Persistence Modules. arXiv pre-print 2010.15623, 2020

[18] Hatcher, A. Algebraic Topology. Cambridge Univ. Press, 2001.

[19] Hu, X., Fuxin, L., Samaras, D., Chen, C., Topology-Preserving Deep Image Segmentation, 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

[20] Lesnick, M. and Wright, M. Interactive Visualization of 2-D Persistence Modules. arXiv pre-print 1512.00180, 2015

[21] Nicolau, M., Levine, A. and Carlsson, G., Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. PNAS April 26, 2011 108 (17) 7265-7270

[22] Varadarajan, V. Weak convergence of measures on separable metric spaces. Sankhyā: The Indian Journal of Statistics, Vol. 19, No. 1/2 (Feb., 1958), pp. 15-22

[23] Vipond, O. Multiparameter persistence landscapes. Journal of Machine Learning Research, 21(61):1–38, 2020.

[24] Wu Z., Song S., Khosla A., Yu F., Zhang L., Tang X. and Xiao, J. 3D ShapeNets: A Deep Representation for Volumetric Shapes Proceedings of 28th IEEE Conference on Computer Vision and Pattern Recognition, 2015

[25] Zomorodian A. and Carlsson, G. Computing persistent homology. Discrete & Computational Geometry, 33(2):249-274, 2005.

A SINGLE-PARAMETER PERSISTENCE HOMOLOGY

Let M be a compact metric space with distance function $d_M : M \times M \to \mathbb{R}$. In practice, M will usually be a finite subset of Euclidean space $\mathbb{R}^n$, but the results here are stated in more generality. The definition of persistence homology uses the language of simplicial complexes. Here we outline the key concepts, and refer the reader to [7] for a detailed treatment and [18] for an introduction to simplicial homology. See also [13] and [25] for the original papers in the setting where M is a finite set.

A simplicial complex $\mathcal{C}$ is a set of simplexes (i.e. points, lines, triangles, and their higher-dimensional counterparts) such that any face of a simplex in $\mathcal{C}$ is also in $\mathcal{C}$, and the intersection of two simplices in $\mathcal{C}$ is either empty or a face of both simplices. The simplicial complexes that are of interest to us are the Vietoris-Rips complexes $\text{Rips}_s(M)$, defined for a metric space M and $s \geq 0$. A simplex $[m_0, \cdots, m_k] \in \text{Rips}_s(M)$ if $d_M(m_i, m_j) \leq s$ for $0 \leq i \leq k$. Note that if $s \leq s'$, then there is an inclusion from $\text{Rips}_s(M)$ to $\text{Rips}_{s'}(M)$; we refer to the family $\{\text{Rips}_s(M)\}_{s\geq 0}$ as a filtered simplicial complex.

The persistent barcode $\beta_i(M)$ of this filtered simplicial complex is obtained by considering the homology groups of the simplicial complexes, $H^i(\text{Rips}_s(M))$. These are vector spaces equipped with linear maps $H^i(\text{Rips}_s(M)) \to H^i(\text{Rips}_{s'}(M))$ coming from the above inclusions. The persistent barcode $\beta_i(M)$ is a statistical summary consisting of intervals $\{[b_k, d_k]\}_{1\leq k\leq n}$. The bottleneck distance d_b between two persistent barcodes is the smallest ϵ such that there is an ϵ-matching between the two barcodes (i.e. any interval $[b_k, d_k]$ from one barcode of length more than ϵ must be matched to an interval $[b'_j, d'_j]$ from the other barcode, with the Cartesian distance between the two points in $\mathbb{R}^2$ being less than ϵ).

B CONVERGENCE OF THE FIBERED BARCODE FOR TWO-PARAMETER PERSISTENCE HOMOLOGY

B.1 STABILITY RESULTS FOR PERSISTENCE HOMOLOGY

One of the most important properties of persistence barcodes is their stability with respect to the Haussdorff distance, which was established in [11] and [7]. Given two subsets M_1, M_2 of a compact metric space M, recall that the Haussdorff distance:

$$d_H(M_1, M_2) = \max\{ \sup_{m_1\in M_1} \inf_{m_2\in M_2} d(m_1, m_2), \\ \sup_{m_2\in M_2} \inf_{m_1\in M_1} d(m_1, m_2)\}$$

The stability property is the following:

$$d_b(\beta_i(M_1), \beta_i(M_2)) \leq 2d_H(M_1, M_2)$$

This statement is often phrased using the Gromov-Haussdorff distance instead, but we will not need that level of generality here. Now we state extensions of these results to the multi-parameter setting, specifically for the multi-level set filtration and the Degree-Rips filtration.

Definition B.1. *Let $\widehat{f} : M \to M \times \mathbb{R}$ be the embedding defined by $\widehat{f}(m) = (m, f(m))$. Given any subset $S \subset M$, let $S_f \subset M \times \mathbb{R}$ be its image under the map $\widehat{f}$.*

Proposition B.2. *Let M be a compact metric space equipped with a function $f : M \to \mathbb{R}$. Given subsets S, S' of M:*

$$d_M(\beta_f(S), \beta_f(S')) \leq 2d_H(S_f, S'_f)$$

See Appendix A for a more detailed statement and proof.

B.2 Multiparameter sublevel-set filtrations

Let M be a compact metric space, and $f : M \to \mathbb{R}$ is a continuous function. Denote by $\beta_f(M)$ its fibered barcode, as defined in Section **??**. Suppose that we observe a sample $\mathbb{X} = \{X_1, \cdots, X_n\}$ drawn from an unknown measure μ on M, and the values $f(X_i)$ for $1 \leq i \leq n$. In this section we analyze the performance of the quantity $\beta_f(\mathbb{X})$ as an estimator of $\beta_f(M)$. First we extend Theorem 3.1 from [15] to this setting.

Theorem B.3. *Let $X_1, X_2, \cdots$ be i.i.d valued random variables chosen from the measure μ on M, which is supported on a compact subset $\mathbb{X}_\mu$. Let $\mathbb{X}_n = \{X_1, \cdots, X_n\}$. The following holds almost surely:*

$$\beta_f(\mathbb{X}_\mu) = \lim_{n \to \infty} \beta_f(\mathbb{X}_n)$$

Next we study the convergence rate for the above result with respect to the matching distance on fibered barcodes, extending Corollary 3 of [9]. Let $M_f \subset M \times \mathbb{R}$ be as defined in Definition B.1, and $i_f : M \to M_f$ be the natural map. Let $\mu_f = i_{f*}\mu$ be the pushforward measure on M_f.

Assumption 1. Suppose that the induced measure μ_f satisfies the (a, b)-standard assumption, for some fixed constants $a, b > 0$: given any $x \in M \times \mathbb{R}$ and $r > 0$, $\mu(B(x, r)) \geq \min\{ar^b, 1\}$.

Theorem B.4. *Suppose that we have a sample of n points $\mathbb{X}_n = \{X_1, \cdots, X_n\}$ drawn from an unknown measure μ.*

$$\mathbb{P}(d_M(\beta_f(\mathbb{X}_\mu), \beta_f(\mathbb{X}_n)) > \epsilon) < \frac{2^b}{a\epsilon^b} exp(-na\epsilon^b)$$

The above theorem can be used to construct confidence sets for the fibered barcode, following Section 3.4 in [9].

Proposition B.5. *Let $\alpha \in (0, 1)$ be chosen arbitrarily. Suppose c satisfies $\frac{e^{-c}}{c} < \frac{\alpha}{n2^b}$. The following is a confidence interval for $\beta_f(\mathbb{X}_\mu)$ with level $1 - \alpha$:*

$$B_{d_M}\left(\beta(\mathbb{X}_n), \sqrt[b]{\frac{c}{na}}\right)$$

B.3 Risk analysis of the closest subsample for multi-level set filtrations

We return to the setting from Sections B.2. Assume that $M = \mathbb{X}$ is a finite metric space, equipped with a function $f : \mathbb{X} \to \mathbb{R}$ and let μ be the discrete uniform distribution centered on $\mathbb{X}$. In particular, we assume that the induced measure μ_f satisfies the (a, b)-standard assumption. Let $\mathbb{X}_{[1]}, \cdots, \mathbb{X}_{[m]}$ be m independent subsamples of size n chosen from μ.

Definition B.6. *Define the closest subsample as follows (here H denotes the Haussdorff distance between the finite metric spaces $\mathbb{X}$ and $\mathbb{X}_i$, and the notation $\mathbb{X}_{[i],f}$ was introduced in Definition B.1):*

$$\widehat{\mathbb{X}}_n^{(m)} = \underset{1 \leq i \leq m}{arg\ min}\ d_H(\mathbb{X}_f, \mathbb{X}_{[i],f})$$

The following proposition follows from Theorem B.4.

Proposition B.7.

$$\mathbb{P}(d_M(\beta_f(\widehat{\mathbb{X}}_n^{(m)}), \beta_f(\mathbb{X})) > \epsilon) \leq \left[\frac{2^b}{a\epsilon^b} exp(-na\epsilon^b)\right]^m$$

C PROOFS

C.1 STABILITY OF MATCHING DISTANCE FOR MULTI-PARAMETER PERSISTENCE FILTRATIONS

In this subsection we prove Proposition 2.6, and establish that the fibered code $\beta_f(M)$ is well-defined for any compact metric space M (in Section 2.2.1, $\beta_f(M)$ was defined for finite metric spaces M).

First we establish Proposition 2.6 in the setting where M is a finite metric space, keeping the notation from Section 2.2.1. In particular, recall that $d : M \times M \to \mathbb{R}$ is the distance function, $\mathcal{L}$ is the space of all affine lines in $\mathbb{R}^2$ with non-negative slope, and $L \in \mathcal{L}$ is a line $y = l_1 x + l_2$.

Proposition 2.6

$$d_M(\beta_f(S), \beta_f(S')) \leq 2d_H(S_f, S'_f)$$

Proof of Proposition 2.6, when M is finite. By definition of the multiparameter interleaving distance, the above statement is equivalent to the below inequality. Here $\text{Rips}^f_L(S) = \{\text{Rips}^f_{L,a}(S)\}_{a\in\mathbb{R}}$ is the filtered simplicial complex defined via $\text{Rips}^f_{L,a}(S) = \text{Rips}^f_{t_1(a),t_2(a)}(S)$ where $t_1(a) = \frac{a}{\sqrt{l_1^2+1}}, t_2(a) = \frac{l_1 a + l_2}{\sqrt{l_1^2+1}}$.

$$w(L)d_B(\text{Rips}^f_L(S), \text{Rips}^f_L(S')) \leq 2d_H(S_f, S'_f)$$

To prove this, we follow the approach used in Lemma 4.3 of [7] to prove the analogous result in the the one-parameter setting, and start by recalling the definitions.

Definition C.1. *Let $S = (S_a)_{a\in\mathbb{R}}$ and $T = (T_a)_{a\in\mathbb{R}}$ be filtered simplicial complexes with vertex sets X and Y. A multivalued map $C : X \twoheadrightarrow Y$ is ϵ-simplicial if for any $\sigma \in S_a$ and $a \in \mathbb{R}$, every finite subset of $C(\sigma)$ is a simplex of $T_{a+\epsilon}$.*

Proposition 4.2 of [7] states that if $C : X \twoheadrightarrow Y$ is a multi-valued map with an inverse C^T, then if C and C^T are both ϵ-simplicial, then the persistence modules $H(S)$ and $H(T)$ are ϵ-interleaved. Now it suffices to show that the filtered complexes $\text{Rips}^f_L(S)$ and $\text{Rips}^f_L(S')$ are ϵ-interleaved, when $\epsilon > \frac{2}{w(l)} d_H(S_f, S'_f)$. Consider the subset $X \subset S \times S'$ of points (u, u') such that $d(u, u') \leq d_H(S_f, S'_f)$, and let C be the corresponding multi-valued map.

Given a simplex $\sigma \in \text{Rips}^f_L(S)_a$, we must show that any finite subset $\tau \in C(\sigma)$ lies in $\text{Rips}^f_L(S')_{a+\epsilon}$. Given any two points $u'_1 = (m'_1, f'_1), u'_2 = (m'_2, f'_2) \in C(\sigma)$, suppose $u'_1 \in C(u_1), u'_2 \in C(u_2)$ for $u_1 = (m_1, f_1), u_2 = (m_2, f_2) \in \sigma$. Then for $i = 1, 2$:

$$\begin{aligned}
d(m_1, m_2) \leq \frac{a}{\sqrt{l_1^2+1}}; \qquad f_i &\leq \frac{l_1 a + l_2}{\sqrt{l_1^2+1}} \\
d(m'_1, m'_2) &\leq d(m_1, m_2) + d(m_1, m'_1) + d(m'_2, m_2) \\
&\leq \frac{a}{\sqrt{l_1^2+1}} + 2d_H(U, U') \leq \frac{a}{\sqrt{l_1^2+1}} + w(l)\epsilon \\
&\leq \frac{a+\epsilon}{\sqrt{l_1^2+1}} \\
f'_i \leq \frac{l_1 a + l_2}{\sqrt{l_1^2+1}} + |f'_i - f_i| &\leq \frac{l_1 a + l_2}{\sqrt{l_1^2+1}} + \epsilon w(L) \\
&\leq \frac{l_1 a + l_2}{\sqrt{l_1^2+1}} + \epsilon \frac{l_1}{\sqrt{l_1^2+1}} = \frac{l_1(a+\epsilon) + l_2}{\sqrt{l_1^2+1}}
\end{aligned}$$

It now follows that $\tau \in C(\sigma)$ lies in $\text{Rips}^l(U)_{a+\epsilon}$, completing the proof. □

For the rest of this section, M will be a compact metric space. Now we will show that $\beta_f(M)$ can be defined in this level of generality, and that Proposition 2.6 holds in this setting. Both of these follow from Proposition C.4 below, which we will establish following the approach in Section 2.4 of [15].

Definition C.2. *Let $F(M)$ denote the set of finite non-empty subsets of the metric space M. Let $K(M)$ denote the set of compact subsets of the metric space M.*

Definition C.3. *Denote by $\mathcal{B}^f_\infty$ the space of all fibered barcodes, viewed as a metric space with the distance given by the matching distance between two fibered barcodes. Let $\widehat{\mathcal{B}}^f_\infty$ be the completion of this metric space.*

Proposition C.4. *There is a unique continuous extension $K(M) \to \widehat{\mathcal{B}}^f_\infty$ of the map $\beta_f : F(M) \to \mathcal{B}^f_\infty$, which we denote by the same symbol. This extended map is also Lipschitz continuous with Lipschitz constant* 2*.*

Proof. Given a compact subset $K \subset M$, let $\mathbb{X} = \{X_1, X_2, \cdots\}$ be i.i.d random variables chosen from the uniform distribution on K, and for each positive integer n let $\mathbb{X}_n = \{X_1, X_2, \cdots, X_n\}$. Given $\epsilon > 0$, it is well-known that the following statements hold almost surely: for $n \gg 0$, $d_H(K, \mathbb{X}_n) < \epsilon$ (see Lemma 5.1 above for the statement and Lemma 3.2 of [15] for the proof). It then follows from that the sequence $\{\beta_f(\mathbb{X}_n)\}_{n\in\mathbb{Z}}$ converges for a suitable choice of $\mathbb{X}$, so $\beta_f(K)$ can be defined as its limit. It is easy to check that it does not depend on the choice of $\mathbb{X}$, and that Lipschitz continuity is satisfied. □

C.2 Proof of key results

Now we prove the key results in Sections 3 and 4.

Proof of Theorem B.3. This follows immediately from Proposition B.2 combined with Lemma 3.2 in [15]. □

Proof of Theorem B.4. This follows from Proposition B.2, combined with Theorem 2 of [9]. Recall that the notation $\mathbb{X}_{\mu,f}$ was introduced in Definition B.1.

$$\begin{aligned}\mathbb{P}(d_M(\beta_f(\mathbb{X}_\mu), \beta_f(\mathbb{X}_n)) > \epsilon) &< \mathbb{P}(d_H(\mathbb{X}_{\mu,f}, \mathbb{X}_{n,f}) > 2\epsilon) \\ &< \frac{2^b}{a\epsilon^b}\exp(-na\epsilon^b)\end{aligned}$$

□

Proof of Theorem B.5. This follows immediately from Theorem B.4. If $\epsilon = \sqrt[b]{\frac{c}{na}}$, then $na\epsilon^b = c$ and the conclusion follows:

$$d_M(\beta_f(\mathbb{X}_\mu), \beta_f(\mathbb{X}_n)) > \epsilon) < \frac{2^b}{a\epsilon^b}\exp(-na\epsilon^b) < \alpha$$

□

Proof of Theorem B.7. This follows from Proposition B.4, combined with Proposition B.2 above:

$$\begin{aligned}\mathbb{P}(d_M(\beta_f(\widehat{\mathbb{X}}^{(m)}_n), \beta_f(\mathbb{X})) > \epsilon) &\leq \mathbb{P}(d_H(\widehat{\mathbb{X}}^{(m)}_{n,f}, \mathbb{X}_f)) > 2\epsilon) \\ &= \mathbb{P}(d_H(\mathbb{X}_{[1],f}, \mathbb{X}_f)) > 2\epsilon)^m \\ &\leq \left[\frac{2^b}{a\epsilon^b}\exp(-na\epsilon^b)\right]^m\end{aligned}$$

□