Finding the Homology of Decision Boundaries with Active Learning

Weizhi Li  
Arizona State University  
weizhili@asu.edu

Gautam Dasarathy  
Arizona State University  
gautamd@asu.edu

Karthikeyan Natesan Ramamurthy  
IBM Research  
knatesa@us.ibm.com

Visar Berisha  
Arizona State University  
visar@asu.edu

Abstract

Accurately and efficiently characterizing the decision boundary of classifiers is important for problems related to model selection and meta-learning. Inspired by topological data analysis, the characterization of decision boundaries using their homology has recently emerged as a general and powerful tool. In this paper, we propose an active learning algorithm to recover the homology of decision boundaries. Our algorithm sequentially and adaptively selects which samples it requires the labels of. We theoretically analyze the proposed framework and show that the query complexity of our active learning algorithm depends naturally on the intrinsic complexity of the underlying manifold. We demonstrate the effectiveness of our framework in selecting best-performing machine learning models for datasets just using their respective homological summaries. Experiments on several standard datasets show the sample complexity improvement in recovering the homology and demonstrate the practical utility of the framework for model selection.

1 Introduction

Meta learning refers to a family of algorithms that learn to generalize the experience from learning a task and therefore dubbed “learning to learn” [1]. Among many of these experience, the complexity of the target data is considered important in the classification task [2]. A common characterization of data complexity is to understand the geometry of the decision boundary; for example, by using topological data analysis (TDA) [3, 4, 5]. This scenario makes sense in settings where large corpora of labeled training data are available to recover the persistent homology of the decision boundary for use in downstream machine learning tasks [5, 6, 7, 8]. However the utility of this family of methods is limited in applications where labeled data is expensive to acquire.

In this paper, we explore the intersection of active learning and topological data analysis for the purposes of efficiently learning the persistent homology of the decision boundary in classification problems. In contrast to the standard paradigm, in active learning, the learner has access to unlabeled data and sequentially selects a set of points for an oracle to label. We propose an efficient active learning framework that adaptively select points for labeling near the decision boundary. A theoretical analysis of the algorithm results in an upper bound on the number of samples required to recover the decision boundary homology. Naturally, this query complexity depends on the intrinsic complexity of the underlying manifold.

There have been several other studies that have explored the use of topological data analysis to characterize the decision boundary in classification problems. In [9], the authors use the persistent
homology of the decision boundary to tune hyperparameters in kernel-based learning algorithms. They later extended this work and derived the conditions required to recover the homology of the decision from only samples [5]. Other works have explored the use of other topological features to characterize the difficulty of classification problems [10][11][7]. While all previous work assumes full knowledge of data labels, only samples near the decision boundary are used to construct topological features. We directly address this problem in our work by proposing an active approach that adaptively and sequentially labels only samples near the decision boundary, thereby resulting in significantly reduced query complexity. To the best of our knowledge, this is the first work that explores the intersection of active learning and topological data analysis.

Our main contributions are as follows:

- We introduce a new algorithm for actively selecting samples to label in service of finding the persistent homology of the decision boundary. We provide theoretical conditions on the query complexity that lead to the successful recovery of the decision boundary homology.
- We evaluate the proposed algorithm for active homology estimation using synthetic data and compare its performance to a passive approach that samples data uniformly. In addition, we demonstrate the utility of our approach relative to a passive approach on a stylized model selection problem using real data.

## 2 Preliminaries

In this section, we define the decision boundary manifold and discuss the labeled Čech Complex [5] which we then use to estimate the homology of this manifold from labeled data. For more background and details, we direct the reader to the appendix.

### 2.1 The Decision Boundary Manifold and Data

Let $\mathcal{X}$ be a Euclidean space that denotes the domain/feature space of our learning problem and let $\mu$ denote the standard Lebesgue measure on $\mathcal{X}$. We will consider the binary classification setting in this paper and let $Y = \{0, 1\}$ denote the label set. Let $p_{XY}$ denote a joint distribution on $\mathcal{X} \times Y$. Of particular interest to us in this paper is the so-called Bayes decision boundary $\mathcal{M} = \{x \in \mathcal{X} | p_{Y|X}(1|x) = p_{Y|X}(0|x)\}$. Indeed, identifying $\mathcal{M}$ is equivalent to being able to construct the provably optimal binary classifier called the Bayes optimal predictor:

$$f(x) = \begin{cases} 1 & \text{if } p_{Y|X}(1 | x) \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Following along the lines of [5], the premise of this paper relies on supposing that the set $\mathcal{M}$ is in fact a reasonably well-behaved manifold\footnote{Note that it is conceivable that the decision boundary is not strictly a manifold. While this assumption is critical to the rest of this paper, it is possible to extend the results here by following the theory in [12]. We will leave a thorough exploration of this for future work.}. That is, we will make the following assumption.

**Assumption 1.** The decision boundary manifold $\mathcal{M}$ has a condition number $1/\tau$. The condition number $\frac{1}{\tau}$ is an intrinsic property of $\mathcal{M}$ (assumed to be a submanifold of $\mathbb{R}^d$) and encodes both the local and global curvature of the manifold. The value $\tau$ is the largest number such that the open normal bundle about $\mathcal{M}$ of radius $r$ is embedded in $\mathbb{R}^N$ for every $r < \tau$. E.g., in Figure 1 where $\mathcal{M}$ is a circle in $\mathbb{R}^2$ and $\tau$ is its radius. We refer the reader to the appendix (or [13]) for a formal definition.

Now we will suppose that we have access to $N$ i.i.d samples $\mathcal{D} = \{x_1, ..., x_N\} \subset \mathcal{X}$ that are drawn according to the marginal distribution $p_X$. Notice that in a typical passive learning setting, we assume access to $N$ i.i.d samples from the joint distribution $p_{XY}$. The goal of this paper is to demonstrate that we may obtain labels for far fewer labels than $N$ while achieving similar performance to the passive learning setting if we are allowed to sequentially and adaptively choose the labels observed. Based on the observed data, we define the set $\mathcal{D}_0 = \{x \in \mathcal{D} : f(x) = 0\}$, that is the set of all samples with Bayes optimal label of 0; similarly, we let $\mathcal{D}_1 = \{x \in \mathcal{D} : f(x) = 1\}$.
We now provide a definition of this complex, letting \( \text{Tub}_w(M) \) denote the smallest tubular neighborhood enclosing \( M \). A stylized example that

![Figure 1: An example of \((\epsilon, \gamma)\)-labeled Čech complex, constructed in a tubular neighborhood \( \text{Tub}_{w+\gamma}(M) \) of radius \( w + \gamma \), for a manifold \( M \) of condition number \( 1/\tau \). The overlap between the two classes (\( \mathcal{D} \)) is contained in \( \text{Tub}_{w}(M) \). The complex is constructed on samples in class 0, by placing balls of radius \( \epsilon \) (\( B_{\epsilon}(x_i) \)), and is “witnessed” by samples in class 1. \( \mathcal{X} \) is the compact probability space for the data. Each triangle is assumed to be a 2-simplex in the simplicial complex. Note that we keep the samples from both classes sparse for aesthetic reasons.

2.2 The Labeled Čech Complex

As outlined in Section 1, our goal is to recover the homological summaries of \( M \) from data. Homological summaries such as Betti numbers estimate the number of connected components and the number of holes of various dimensions that are present in \( M \). Since we only have a sample of data points in practice, we first construct a simplicial complex from these points that mimics the shape of \( M \). We can then estimate the rank of any homology group \( H_i \) of dimension \( i \) from this complex. This rank is called the Betti number \( \beta_i \), and informally denotes the number of holes of dimension \( i \) in the complex. The multi-scale estimation of Betti numbers, which involves gradual “thickening” of the complex, results in a persistence diagram \( PD_i \) that encodes the birth and death time of the \( i \)-dimensional holes in the complex. For more background we refer the reader to \([14]\).

In \([5]\), the authors consider the passive learning setting for estimating the homology of \( M \) and propose a simplicial approximation for the decision boundary called the Labeled Čech (LC) Complex. We now provide a definition of this complex, letting \( B_{\epsilon}(x) \) denote a ball of radius \( \epsilon \) around \( x \). We refer the reader to the appendix or \([5]\) for the more details.

**Definition 1.** Given \( \epsilon, \gamma > 0 \), an \((\epsilon, \gamma)\)-labeled Čech complex is a simplicial complex constructed from a collection of simplices such that each simplex \( \sigma \) is formed on the points in a set \( S \subseteq D^0 \) witnessed by the reference set \( D^1 \) satisfying the following conditions: (a) \( \bigcap_{x_i \in \sigma} B_{\epsilon}(x_i) \neq \emptyset \), where \( x_i \in S \) are the vertices of \( \sigma \). (b) \( \forall x_i \in S \subseteq D^0, \exists x_j \in D^1 \) such that, \( \|x_i - x_j\|_2 \leq \gamma \).

The set \( S \) is used to construct the LC complex witnessed by the reference set \( D^1 \). This allows us to infer that each vertex of the simplices \( \sigma \) are within distance \( \gamma \) to some point in \( D^1 \). The authors in \([5]\) show that, under certain assumption on the manifold and the distribution, provided sufficiently many random samples (and their labels) drawn according \( p_{XY} \), the set \( U = \bigcup_{x_i \in \sigma} B_{\epsilon}(x_i) \) forms a cover of \( M \) and deformation retracts to \( M \). Moreover, the nerve of the covering is homotopy equivalent to \( M \). The assumptions under which the above result holds also turns out to be critical to achieve the results of this paper, and hence we will devote the rest of this section to elaborating on these.

Before stating our assumptions, we need a few more definitions. For the distribution \( p_{XY} \), we will let

\[
\mathcal{D} \triangleq \{ x \in \mathcal{X} : p_{X|Y}(x \mid 1)p_{X|Y}(x \mid 0) > 0 \}.
\]

In other words, \( \mathcal{D} \) denotes the region of the feature space where both classes overlap, i.e., both class conditional distributions \( p_{X|Y}(\cdot \mid 1) \) and \( p_{X|Y}(\cdot \mid 0) \) have non-zero mass. For any \( r > 0 \), we let \( \text{Tub}_r(M) \) denote a “tubular” neighborhood of radius \( r \) around \( M \) \([13]\). Furthermore, we write \( \text{Tub}_w(M) \) to denote the smallest tubular neighborhood enclosing \( \mathcal{D} \). A stylized example that
highlights the relationship between these parameters is shown in Figure 1. In the sequel, we will introduce the related assumption and the lemma underlying the results of this paper (similar to those in [13, 5]). All assumptions and follow-on results are dependent on two parameters that are specific to the joint distribution: \( w \) - the amount of overlap between the distributions and \( \tau \) - the global geometric property of the decision boundary manifold.

**Assumption 2.** \( w < (\sqrt{\delta} - \sqrt{8})\tau \).

A principal difference between [13] and our work is that [13] has direct access to the manifold and supposes that all generated samples are contained within \( \text{Tub}_{(\sqrt{\delta} - \sqrt{8})\tau}(\mathcal{M}) \); this is one of the sufficient conditions for manifold reconstruction from samples. In contrast, as in [5], we do not have direct access to the decision boundary manifold. Rather, certain properties of \( \mathcal{M} \) are inferred from the labels. In this paper, we show that we can infer the homology of \( \mathcal{M} \) with far fewer labels if we are allowed to sequentially and adaptively choose which labels to obtain.

Since the generated samples do not necessarily reside within \( \text{Tub}_{(\sqrt{\delta} - \sqrt{8})\tau}(\mathcal{M}) \), it is not immediately apparent that it is possible to find an \( S \) (see Definition 1) that is entirely contained in \( \text{Tub}_{(\sqrt{\delta} - \sqrt{8})\tau}(\mathcal{M}) \). However, Assumption 2 allows us to guarantee precisely this. To see this, we will first state the following lemma.

**Lemma 1.** Provided \( D^0 \) and \( D^1 \) are both \( \frac{\gamma}{2} \)-dense in \( \mathcal{M} \), then \( S \) is contained in \( \text{Tub}_{w + \gamma}(\mathcal{M}) \) and it is \( \frac{\gamma}{2} \)-dense in \( \mathcal{M} \).

That is, the lemma tells us that as long as we choose our sampling strategy to guarantee that both \( D^0 \) and \( D^1 \) are \( \gamma/2 \)-dense in \( \mathcal{M} \), and provided we choose \( \gamma < (\sqrt{\delta} - \sqrt{8})\tau - w \), we can guarantee that \( S \subseteq \text{Tub}_{(\sqrt{\delta} - \sqrt{8})\tau}(\mathcal{M}) \). \( D^0 \) and \( D^1 \) being \( \frac{\gamma}{2} \)-dense in \( \mathcal{M} \) implies \( \|x_i - x_j\|_2 < \gamma \), where \( x_i \in D^0 \cap B_{\frac{\gamma}{2}}(p) \) and \( x_j \in D^1 \cap B_{\frac{\gamma}{2}}(p) \) for every \( p \in \mathcal{M} \).

Therefore, as the distance from \( x_i \in S \subseteq D^0 \) to \( x_j \in D^1 \) is bounded by \( \gamma \) (see Definition 1(b)), \( S \) is also \( \frac{\gamma}{2} \)-dense in \( \mathcal{M} \). This immediately implies \( S \) is \( (w + \gamma) \)-dense in \( \mathcal{M} \). Therefore, with \( S \subseteq \text{Tub}_{(\sqrt{\delta} - \sqrt{8})\tau}(\mathcal{M}) \), \( S \) being \( (w + \gamma) \)-dense in \( \mathcal{M} \) and an appropriate \( \epsilon \) properly selected, we will have the \((\epsilon, \gamma)\)-LC complex homotopy equivalent to \( \mathcal{M} \) per theorems in [5, 13]. We now state the following proposition.

**Proposition 1.** \((\epsilon, \gamma)\)-LC complex is homotopy equivalent to \( \mathcal{M} \) as long as (1) \( \gamma < (\sqrt{\delta} - \sqrt{8})\tau - w \); (2) \( D^0 \) and \( D^1 \) are \( \frac{\gamma}{2} \)-dense in \( \mathcal{M} \) and (3) \( \epsilon \in (\frac{w + \gamma + \tau - \sqrt{(w + \gamma)^2 + \tau^2 - 6\tau (w + \gamma)}}{2}, \frac{(w + \gamma + \tau) + \sqrt{(w + \gamma)^2 + \tau^2 - 6\tau (w + \gamma)}}{2}) \).

A pictorial description of relations between \( w, \gamma \) and \( \tau \) is in Figure 1. In the stylized example in Figure 1, \( \mathcal{M} \) is a circle, \( \text{Tub}_{w + \gamma}(\mathcal{M}) \) is an annulus and the radius \( \epsilon \) of the covering ball \( B_{\epsilon}(x) \) is constrained by \( \tau \).

Figure 2: The proposed active learning framework for finding the homology of decision boundaries.

## 3 Active Learning for Finding the Homology of Decision Boundaries

As the definitions above and results from [5] make clear, constructing a useful \( \text{LC} \) complex requires sampling both class-conditional distributions in the region around the decision boundary to a sufficient resolution. The key insight of our paper is to devise a framework based on active learning that sequentially and adaptively decides where to obtain data and therefore query-efficiently samples points near the decision boundary. In what follows, we will provide a brief description of our algorithm, and then we will establish rigorous theoretical guarantees on the query complexity of the proposed algorithm.

A set \( W \subseteq \mathcal{D} \) is \( \frac{\gamma}{2} \)-dense in \( \mathcal{M} \) if for every \( p \in \mathcal{M} \) there exists a \( x \in W \) such that \( \|p - x\|_2 < \frac{\gamma}{2} \). In other words, there exists at least one \( x \in W \) in \( B_{\frac{\gamma}{2}}(p) \) for every \( p \in \mathcal{M} \).
3.1 The Active Learning Algorithm

A schematic diagram of the proposed active learning framework is presented in Figure 2. As illustrated, the framework starts from sampling sufficient unlabeled data from \( p_X \). Subsequently, the framework takes as input an unlabeled dataset \( \mathcal{D} \), and this dataset is used to generate an appropriate graph on the data. This graph is then used to iteratively query labels near the decision boundary. The subset of labeled samples are used to estimate the homology, resulting in the persistence diagram of the LC complex. We briefly outline the label query and homology estimation phases below, and refer the reader to the supplementary material for the full details.

**Label query phase:** The label query phase starts with constructing a graph \( G = (\mathcal{D}, E) \) from the unlabeled dataset \( \mathcal{D} \). While other choices are possible, we will suppose that the graph we construct is either a \( k \)-radius near neighbor or a \( k \)-nearest neighbors graph. After graph construction, a graph-based active learning algorithm \((S^2) \) path [15] accepts the graph \( G = (\mathcal{D}, E) \) and selects the data points whose labels it would like to see. This selection is based on the structure of the graph and all previous gathered labels. Specifically, \( S^2 \) continually queries for the label of the vertex that bisects the shortest path between any pair of oppositely-labeled vertices. The authors in [15] show that \( S^2 \) provably query and efficiently locate the cut-set in this graph (i.e., the edges of the graph that have oppositely labeled vertices). As a result, the query phase outputs a set \( \hat{\mathcal{D}} \) associated with the labels that is near the decision boundary.

**Homology estimation phase:** During the homology estimation stage, we construct an approximation of the LC complex from the query set \( \hat{\mathcal{D}} \). Specifically, we construct the locally scaled labeled Vietoris-Rips (LS-LVR) complex introduced in [5]. Sticking to the query set \( \hat{\mathcal{D}} \) as an example, there are two steps to construct the LS-LVR complex: (1) Generate an initial graph from \( \hat{\mathcal{D}} \) by creating an edge set \( \hat{E} \) as follows: \( \hat{E} = \{(x_i, x_j) \mid (x_i, x_j) \in \mathcal{D}^2 \land y_i \neq y_j \land \|x_i - x_j\| \leq \kappa \sqrt{\rho_i \rho_j} \} \). Here, \( \kappa \) is a scale parameter, \( \rho_i \) is the smallest radius of a sphere centered at \( x_i \) to enclose \( k \)-nearest opposite class neighbors and \( \rho_j \) has a similar definition. This creates a bipartite graph where every edge connect points of opposite class; (2) Connect all 2-hop neighbors to build a simplicial complex. Varying scale parameter \( \kappa \) produces a filtration of the simplicial complex and generates the persistent homology such as the persistent diagrams.

3.2 Query Complexity of the Active Learning Algorithm

Let \( G = (\mathcal{D}, E) \) denote a \( k \)-radius neighbor graph constructed from the dataset \( \mathcal{D} \). This allows us to define the cut-set \( C = \{(x_i, x_j) \mid y_i \neq y_j \land (x_i, x_j) \in E \} \) and cut-boundary \( \partial C = \{x \in \mathcal{V} : \exists e \in C \text{ with } x \in e \} \). We begin by sketching a structural lemma about the graph \( G \) and refer the reader to the supplementary materials for a full statement and proof.

**Lemma 2.** Suppose \( \mathcal{D}^0 \) and \( \mathcal{D}^1 \) are \( \frac{\gamma}{2} \)-dense in \( \mathcal{M} \), then the graph \( G = (\mathcal{D}, E) \) constructed from \( \mathcal{D} \) is such that \( \mathcal{D}^0 \cap \partial C \) and \( \mathcal{D}^1 \cap \partial C \) are both \( \frac{\gamma}{2} \)-dense in \( \mathcal{M} \) and \( \partial C \subseteq \text{Tub}_{w+\gamma}(\mathcal{M}) \) for \( k = \gamma \).

\( \mathcal{D}^0 \) and \( \mathcal{D}^1 \) being \( \frac{\gamma}{2} \)-dense in \( \mathcal{M} \) indicates the longest distance between \( x_i \in \mathcal{D}^0 \cap B_{\frac{\gamma}{2}}(p) \) and \( x_j \in \mathcal{D}^0 \cap B_{\frac{\gamma}{2}}(p) \) for \( p \in \mathcal{M} \) is \( \gamma \). Therefore, letting \( k = \gamma \) as Lemma 2 suggested results in both \( \mathcal{D}^0 \cap \partial C \) and \( \mathcal{D}^1 \cap \partial C \) being \( \frac{\gamma}{2} \)-dense in \( \mathcal{M} \). Similar to Lemma 1 constructing a graph with a \( \gamma \) radius inevitably results in a subset of points in \( \partial C \) leaking out of \( \text{Tub}_{w}(\mathcal{M}) \) and we formally have \( \partial C \subseteq \text{Tub}_{w+\gamma}(\mathcal{M}) \). The key intuition behind our approach is that \( S^2 \) is naturally turned to focusing the labels acquired within \( \text{Tub}_{w+\gamma}(\mathcal{M}) \). As we show below, this is done in a remarkably query efficient manner, and furthermore, when we obtain labeled data via querying we can construct an LC complex; this allows us to find the homology of the manifold \( \mathcal{M} \). We next need some structural assumptions about the manifold \( \mathcal{M} \).

**Assumption 3.** (a) \( \inf_{x \in \mathcal{M}} \mu_x(B_{\gamma/4}(x)) > \rho_{\gamma/4} \), \( y \in \{0, 1\} \). (b) \( \sup_{x \in \mathcal{M}} \mu_x(B_{(w+\gamma)(x)}) < h_{(w+\gamma)} \). (c) \( \mu_x(\text{Tub}_{w+\gamma}(\mathcal{M})) \leq N_{w+\gamma} h_{w+\gamma} \).

Assumption 3(a) ensures sufficient mass in both classes such that \( \mathcal{D}^0 \) and \( \mathcal{D}^1 \) are \( \frac{\gamma}{2} \)-dense in \( \mathcal{M} \). Assumption 3(b)(c) upper-bounds the measure of \( \text{Tub}_{w+\gamma}(\mathcal{M}) \). Recall that \( G = (\mathcal{D}, E) \) in Lemma 2.
is a labeled graph; we further write $\beta$ to denote the proportion of the smallest connected component with all the examples identically labeled. We lay out our main theorem as follows

**Theorem 1.** Let $N_{w+\gamma}$ be the covering number of the manifold $\mathcal{M}$. Under Assumptions\textsuperscript{[1]} and\textsuperscript{[3]} for any $\delta > 0$, we have that the $(\epsilon, \gamma)$-LC complex estimated by our framework is homotopy equivalent to $\mathcal{M}$ with probability at least $1 - \delta$ provided

\[
|\hat{D}| > \frac{\log \left\{ 1 / \beta \left( 1 - \sqrt{1 - \delta} \right) \right\}}{\log \left\{ 1 / (1 - \beta) \right\}} + |D| N_{w+\gamma} h_{w+\gamma} (\log_2 |D| + 1)
\]

where

\[
|D| > \max \left\{ \frac{1}{P(y = 0) \rho_{\gamma/4}^{\beta}} \left[ \log (2 N_{\gamma/4}) + \log \left( \frac{1}{1 - \sqrt{1 - \delta}} \right) \right], \frac{1}{P(y = 1) \rho_{\gamma/4}^{\beta}} \left[ \log (2 N_{\gamma/4}) + \log \left( \frac{1}{1 - \sqrt{1 - \delta}} \right) \right] \right\}
\]

**Remark 1.** Theorem\textsuperscript{[7]} demonstrates that our active learning framework has a query complexity of $\mathcal{O}(N N_{w+\gamma} h_{w+\gamma} \log_2 N)$. That is, after $\mathcal{O}(N N_{w+\gamma} h_{w+\gamma} \log_2 N)$ queries at most, a $(\epsilon, \gamma)$-LC complex constructed from the queried examples will be homotopy equivalent to $\mathcal{M}$ with high probability. Notice that the intrinsic complexity of the manifold naturally plays a significant role, and the more complex the manifold the more significant gains the active learning framework has over its passive counterpart (cf. eq. 2). In the supplementary material, we also provide a simple and concrete example that numerically shows the improvement in query complexity associated with our proposed framework relative to its passive counterpart.

**Remark 2.** The results of Theorem\textsuperscript{[7]} can be improved by carrying out a more intricate analysis of the active learning algorithm as in \textsuperscript{[15]}. Indeed, one may also replace the $S^2$ algorithm in our framework with a different graph-based active learning algorithm seamlessly to leverage the properties of that algorithm for active homology estimation of decision boundaries. These, and the relaxation of Assumption\textsuperscript{[2]}, are promising directions for future work.

**Remark 3.** Parameters $w$ and $\tau$ are intrinsic properties of $p_{xy}$ and $\mathcal{M}$ and these properties are fixed to a classification problem. Variables $\gamma$ and $\epsilon$ are algorithm variables and they are bounded as stated in Proposition\textsuperscript{[4]}.

We provide a complete proof of Theorem\textsuperscript{[1]} in the supplementary material. However, we will provide some intuition about the operation of our algorithm, and hence to the proof of the theorem here.

The $S^2$ algorithm is split into two phases: uniform querying of labels and querying via path bisection. The uniform querying serves to find a path connecting vertices of opposite labels. The path bisection phase queries at the mid-point of the shortest path that connects oppositely labeled vertices in the underlying graph. As the authors in \textsuperscript{[15]} show, this endows $S^2$ with the ability to quickly narrow in on the cut-boundary $\partial \mathcal{C}$. The uniform querying phase accounts for the first term in eq. 2 which guarantees that there are sufficient paths to identify $\partial \mathcal{C}$ completely.

During the path bisection phase, we take $(\log_2 |\hat{D}| + 1)$ queries at most (this may be tightened using the techniques in \textsuperscript{[15]}) to find the end point of the cut-edge inside a path; this needs to be done at most $|\partial \mathcal{C}|$ to complete the querying phase. Next, with the Assumption\textsuperscript{[2]} and the Lemma\textsuperscript{[2]} it is guaranteed that $\partial \mathcal{C} \subseteq \text{Tub}_{w+\gamma}(\mathcal{M}) \subseteq \text{Tub}_{(\sqrt{\gamma} - \sqrt{\gamma})}(\mathcal{M})$ with $\gamma$ properly selected following Proposition\textsuperscript{[1]}. Therefore, we may use the measure $N_{w+\gamma} h_{w+\gamma}$ from Assumption\textsuperscript{[5]}(b)(c) to upper-bound $|\partial \mathcal{C}|$ which results in the second term of eq. 2. This naturally ties the query complexity to the manifold complexity via $N_{w+\gamma}$ and $\text{Tub}_{w+\gamma}(\mathcal{M})$. Eq. 2 comes from the necessary condition for the LC complex being homotopy equivalent to $\mathcal{M}$, following along the lines of \textsuperscript{[5]}.

![Figure 3: Visualization of the synthetic data.](image-url)
4 Experimental Results

We compare the homological properties estimated from our active learning algorithm to a passive learning approach on both synthetic data and real data.

In the experiments we use the characteristics of homology group of dimension 1 ($\beta_1$, $PD_1$). We chose to use dimension 1 since [5] shows that this provides the best topological summaries for applications related to model selection.

Using the synthetic data, we study the query complexity of active learning by examining the homological summaries $\beta_1$ and $PD_1$. For real data, we estimate $PD_1$ of the Banknote, MNIST and CIFAR10 and then utilize $PD_1$ to do model selection from several families of classifiers.

4.1 Experiments on Synthetic Data

The synthetic data in Figure 3 has decision boundaries that are homeomorphic to two disjoint circles. This dataset has 2000 examples. Please see the appendix for more details. Clearly from Figure 3 $\beta_1$ of the decision boundary is two.

Per the first step of our active learning algorithm, we construct a $k$-radius NN graph with $k = 0.65$. The scale parameter is set assuming we have full knowledge of the decision boundary manifold. Subsequently, we use $S^2$ to query the labels of examples on the created graph. After the label query phase, we construct the LS-LVR complex with the queried samples and compute $\beta_1$ and $PD_1$ using the Ripser package [16] and its interface [17]. For the passive learning baseline, we uniformly query the examples with all other aspects of the experiment remaining identical to the active case. We also compute $\beta_1$ and $PD_1$ from the complete dataset and consider them as the “ground-truth” homology summaries. We evaluate the similarities between the estimated homology summaries and the ground-truth homology summaries to show the effectiveness of our active learning framework.

We compare the bottleneck distance [18] [19] between the ground-truth and estimated values of $PD_1$ for different percent of data labeled. These results are shown on Figure 4. As is clear from the figure, the bottleneck distance for our active learning framework decreases faster than the passive learning approach and perfectly recovers the homology with only 50% of data. A visualization of the query process is shown on Figure 5. As expected, the active learning framework selects more examples to query near the decision. Please refer to the appendix to evaluate the performance of the active learning framework for different $k$-radius near neighbor graphs and $\beta_1$ recovery.

4.2 Experiments on Real Data

To demonstrate the effectiveness of our active learning framework on real data, we consider the classifier selection problem discussed in [5]. A bank of pretrained classifiers is accessible in the marketplace and customers will select a proper one without changing the hyperparameters of the classifiers. We consider two selection strategies as follow. One is topologically based where a classifier with the smallest bottleneck distance from the $PD_1$ of queried data is selected. The other one is to ensemble the topologically selected classifier and the classifier selected based on the validation error of the queried data. We ensemble these two classifiers by averaging the output probabilities.

![Figure 4: Bottleneck distance from ground-truth $PD_1$ by passive learning and active learning.](image)

![Figure 5: Visualization of the query process by passive learning (top row) and our active learning framework (bottom row). More examples (highlighted by green) near the decision boundaries are selected to query in the proposed framework.](image)
## Table 1: Average test error rates (five trials) on banknote, MNIST and CIFAR10 for the model selected with 15% pool data. Passive/Active (ens) stands for the classifiers ensembled from two classifiers: one is selected by the homological similarities and the other one is selected by the validation error. The subscript 1, 2 and 3 of the similarities. Passive/Active (ens) stands for the classifiers ensembled from two classifiers: one is selected by the validation error and computing the PD of the decision boundary of the classifier, we simply use the test set input and the classifier output. For the non-ensemble active learning indicates the used 3NN, 5NN and 7NN graphs. Best performance in the non-ensemble and ensemble cases are boldfaced.

<table>
<thead>
<tr>
<th></th>
<th>KNN</th>
<th>SVM</th>
<th>Neural network</th>
<th>Decision tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passive</td>
<td>0.1072 ± 0.0000</td>
<td>0.3753 ± 0.0005</td>
<td>0.4316 ± 0.0000</td>
<td>0.1997 ± 0.0000</td>
</tr>
<tr>
<td>Active¹</td>
<td>0.0783 ± 0.0014</td>
<td>0.3231 ± 0.0012</td>
<td>0.4316 ± 0.0000</td>
<td>0.1901 ± 0.0004</td>
</tr>
<tr>
<td>Active²</td>
<td>0.1017 ± 0.0001</td>
<td>0.3431 ± 0.0012</td>
<td>0.3730 ± 0.0138</td>
<td>0.1744 ± 0.0026</td>
</tr>
<tr>
<td>Active³</td>
<td>0.0346 ± 0.0013</td>
<td>0.0836 ± 0.0133</td>
<td>0.1058 ± 0.0265</td>
<td>0.1613 ± 0.0004</td>
</tr>
<tr>
<td>Passive (ens)</td>
<td>0.0176 ± 0.0000</td>
<td>0.0259 ± 0.0000</td>
<td>0.0068 ± 0.0000</td>
<td>0.0741 ± 0.0000</td>
</tr>
<tr>
<td>Active¹ (ens)</td>
<td>0.0173 ± 0.0000</td>
<td>0.0259 ± 0.0000</td>
<td>0.0039 ± 0.0000</td>
<td>0.0731 ± 0.0000</td>
</tr>
<tr>
<td>Active² (ens)</td>
<td>0.0149 ± 0.0000</td>
<td>0.0259 ± 0.0000</td>
<td>0.0134 ± 0.0001</td>
<td>0.0731 ± 0.0000</td>
</tr>
<tr>
<td>Active³ (ens)</td>
<td>0.0149 ± 0.0000</td>
<td>0.0259 ± 0.0000</td>
<td>0.0072 ± 0.0000</td>
<td>0.0770 ± 0.0000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>KNN</th>
<th>SVM</th>
<th>Neural network</th>
<th>Decision tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passive</td>
<td>0.0129 ± 0.0000</td>
<td>0.0141 ± 0.0000</td>
<td>0.0202 ± 0.0000</td>
<td>0.0332 ± 0.0000</td>
</tr>
<tr>
<td>Active¹</td>
<td>0.0128 ± 0.0000</td>
<td>0.0161 ± 0.0001</td>
<td>0.0150 ± 0.0000</td>
<td>0.0388 ± 0.0001</td>
</tr>
<tr>
<td>Active²</td>
<td>0.0122 ± 0.0000</td>
<td>0.0162 ± 0.0001</td>
<td>0.0177 ± 0.0000</td>
<td>0.0332 ± 0.0000</td>
</tr>
<tr>
<td>Active³</td>
<td>0.0104 ± 0.0000</td>
<td>0.0156 ± 0.0001</td>
<td>0.0388 ± 0.0002</td>
<td>0.0332 ± 0.0000</td>
</tr>
<tr>
<td>Passive (ens)</td>
<td>0.0119 ± 0.0000</td>
<td>0.0124 ± 0.0000</td>
<td>0.0104 ± 0.0000</td>
<td>0.0290 ± 0.0000</td>
</tr>
<tr>
<td>Active¹ (ens)</td>
<td>0.0123 ± 0.0000</td>
<td>0.0119 ± 0.0000</td>
<td>0.0104 ± 0.0000</td>
<td>0.0284 ± 0.0000</td>
</tr>
<tr>
<td>Active² (ens)</td>
<td>0.0108 ± 0.0000</td>
<td>0.0119 ± 0.0000</td>
<td>0.0125 ± 0.0000</td>
<td>0.0284 ± 0.0000</td>
</tr>
<tr>
<td>Active³ (ens)</td>
<td>0.0104 ± 0.0000</td>
<td>0.0119 ± 0.0000</td>
<td>0.0127 ± 0.0000</td>
<td>0.0274 ± 0.0000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>KNN</th>
<th>SVM</th>
<th>Neural network</th>
<th>Decision tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passive</td>
<td>0.3065 ± 0.0002</td>
<td>0.4683 ± 0.0000</td>
<td>0.3185 ± 0.0000</td>
<td>0.3625 ± 0.0000</td>
</tr>
<tr>
<td>Active¹</td>
<td>0.3201 ± 0.0000</td>
<td>0.4591 ± 0.0005</td>
<td>0.3058 ± 0.0006</td>
<td>0.3625 ± 0.0000</td>
</tr>
<tr>
<td>Active²</td>
<td>0.3095 ± 0.0001</td>
<td>0.4007 ± 0.0038</td>
<td>0.3058 ± 0.0006</td>
<td>0.3625 ± 0.0000</td>
</tr>
<tr>
<td>Active³</td>
<td>0.3109 ± 0.0001</td>
<td>0.4464 ± 0.0005</td>
<td>0.3185 ± 0.0000</td>
<td>0.3625 ± 0.0000</td>
</tr>
<tr>
<td>Passive (ens)</td>
<td>0.2987 ± 0.0001</td>
<td>0.2698 ± 0.0000</td>
<td>0.2651 ± 0.0001</td>
<td>0.3137 ± 0.0002</td>
</tr>
<tr>
<td>Active¹ (ens)</td>
<td>0.2911 ± 0.0001</td>
<td>0.2797 ± 0.0000</td>
<td>0.2558 ± 0.0000</td>
<td>0.3146 ± 0.0000</td>
</tr>
<tr>
<td>Active² (ens)</td>
<td>0.2987 ± 0.0001</td>
<td>0.2864 ± 0.0003</td>
<td>0.2649 ± 0.0001</td>
<td>0.3214 ± 0.0005</td>
</tr>
<tr>
<td>Active³ (ens)</td>
<td>0.2935 ± 0.0000</td>
<td>0.2665 ± 0.0000</td>
<td>0.2615 ± 0.0001</td>
<td>0.3221 ± 0.0004</td>
</tr>
</tbody>
</table>

We split the real data to a training set, a test set and lastly unlabeled data pool. The training set is used to generate four different banks of classifiers: KNN with nearest neighbors number ranging from 1 to 29, SVM with polynomial kernel function degree ranging from 1 to 14, decision tree with maximum depth ranging from 1 to 27, and neural networks with layers number ranging from 1 to 6. The test set is used to evaluate the test error of each classifier. The unlabelled data pool is prepared for examples query.

We use the proposed active learning framework to estimate the homological properties of the queried data: constructing a k-nearest neighbors graph, query examples by $S^2$ and computing the PD of the decision boundary of the classifier, we simply use the test set input and the classifier output. Having estimated the homological summaries from the queried data and the classifiers, we compute the bottleneck distance between the PD of the queried data and the classifiers. For the non-ensemble method, we simply select the classifier having the smallest bottleneck distance as a topologically selected classifier. For the ensemble method, we further include an additional classifier selected.
based on the validation error computed from the queried data and ensemble it with the topologically selected classifier.

We implement the above procedure in the datasets Banknote [20], MNIST [21] and CIFAR10 [22]. Banknote contains 1372 instances in two classes with four input features for a binary classification task. We randomly sample 100 examples to construct the training set. Given the small size of the Banknote dataset, we use the rest data as both the test set and unlabelled data pool. Although the test set and the data pool are not rigorously split in the Banknote case, it is still a fair comparison since the performance different is only subject to the query strategy. For the MNIST and the CIFAR10 datasets, we create a digit 1 vs. 8 classification task from the MNIST and an automobile vs. ship classification task from the CIFAR10. We randomly sample the data to create a training set with the size of 200, a test set with the size of 2000, and a unlabelled data pool with the size of 2000.

Table 1 indicates the test error on banknote, MNIST and CIFAR10 for the classifier selected by querying 15% of the unlabelled data pool. We observe that the classifiers selected by our proposed active learning framework generally has a lower test error rate than the passive learning, especially in an ensemble classifier selection framework. As the experimental set-ups are identical except the query strategy, therefore we reasonably attribute the performance improvement to the proposed active learning framework used in the model selection. Figure 6 indicates the performance of the non-ensemble classifiers selected by the homological similarities at the cost of the different proportions of the data pool. As expected, the proposed active learning framework achieves the best model selection faster than the passive learning for all the classifiers families. Note that the selection performance may be unstable with the increasing number of the queries, since the active learning exhausts informative examples rapidly and begin to query noisy examples. In summary, Table 1 and Figure 6 implies the advantage of the active learning in finding good homology summaries is transmitted to the model selection in which the active learning generates a lower error rate performance than the passive learning.

4.3 Homological Properties Analysis on Real Data

We present the homological properties estimated by the passive learning and the proposed active learning framework. Similar to the experiments in the synthetic dataset, we access the whole unlabelled data pool and their labels to compute $\beta_1$ and $PD_1$ and use them as the ground-truth $\beta_1$ and $PD_1$. On the other hand, we query the unlabelled data pool and estimate $\beta_1$ and $PD_1$ from the queried data. As we observe in the Figure 7(a), $\beta_1$ estimated by our active learning algorithm has a more similar trend to the ground-truth $\beta_1$ in all three real datasets. Furthermore, CIFAR10 has a significantly higher $\beta_1$ than MNIST and Banknote datasets indicating more complex decision boundaries. This is consistent with the Table 1 which shows by the error rates that CIFAR10 binary classification task is more difficult than the other two tasks of Banknote and MNIST. Figure 7(b) shows the bottleneck distance between the estimated $PD_1$ and the ground-truth $PD_1$ at the cost of the different proportion of the unlabelled data pool. We observe that the proposed active learning algorithm maintains a smaller bottleneck distance on early stages of query. Such benefits gradually diminish along with the increasing data proportion used for query.

Figure 7: Homological properties for the banknote (left), MNIST (middle) and CIFAR10 (right).

5 Conclusions

We propose an active learning algorithm to find the homology of decision boundaries. We theoretically analyze the query complexity of the proposed algorithm and prove the sufficient conditions to recover the homology of decision boundaries in the active learning setting. The extensive experiments on synthetic and real datasets with the application on model selection corroborate our theoretical results.
Broader Impact

The proposed approach, although has strong algorithmic and theoretical merits, has potential real-world application as we demonstrated.

One of the key uses of this approach is to create efficient summaries of decision boundaries of datasets\[23\] and models. Such summaries can be quite useful in applications like AI model marketplaces\[24\], where data and models can be securely matched without revealing too much information about each other. This is helpful in scenarios where the data is private and models are proprietary or sensitive.

A downside of being able to compute homology of decision boundaries with few examples is that malicious users may be able to learn about the key geometric/topological properties of the models with fewer examples than they would use otherwise. While this in itself may be benign, combined with other methods, they may be able to design better adversarial attacks on this model for instance. Ways of mitigating it in sensitive scenarios include ensuring that users do not issue too many queries of examples close to the boundary successively, since this may be revealing of malicious intent.

Acknowledgements

This work is funded in part by the Office of Naval Research under grant N00014-17-1-2826 and the National Science Foundation under grant OAC-1934766. Karthikeyan Natesan Ramamurthy is employed at IBM Corporation.

References


