Topological Data Analysis of Decision Boundaries with Application to Model Selection

Karthikeyan Natesan Ramamurthy [1] **Kush R. Varshney** [1] **Krishnan Mody** [1 2]

Abstract

We propose the labeled Čech complex, the plain labeled Vietoris-Rips complex, and the locally scaled labeled Vietoris-Rips complex to perform persistent homology inference of decision boundaries in classification tasks. We provide theoretical conditions and analysis for recovering the homology of a decision boundary from samples. Our main objective is quantification of deep neural network complexity to enable matching of datasets to pre-trained models to facilitate the functioning of AI marketplaces; we report results for experiments using MNIST, FashionMNIST, and CIFAR10.

1. Introduction

In supervised learning, the term *model selection* usually refers to the process of using validation data to tune hyperparameters. However, we are moving toward a world in which model selection refers to marketplaces of pre-trained deep learning models in which customers select from a vendor's collection of available models, often without the ability to run validation data through them or being able to change their hyperparameters. Such a marketplace paradigm is sensible because deep learning models have the ability to generalize from one dataset to another (Arpit et al., 2017; Kawaguchi et al., 2017; Zhang et al., 2016). In the case of classifier selection, the use of data and decision boundary complexity measures, such as the critical sample ratio (the density of data points near the decision boundary), can be a helpful tool (Arpit et al., 2017; Ho & Basu, 2002).

In this paper, we propose the use of persistent homology (Edelsbrunner & Harer, 2008), a type of topological data analysis (TDA) (Carlsson, 2009), to quantify the complexity of neural network decision boundaries. Persistent homology involves estimating the number of connected components and the number of holes of various dimensions that are present in the underlying manifold that data samples come from. This complexity quantification can serve multiple purposes, but we focus how it can be used as an aid for matching vendor pre-trained models to customer data. To this end, we must extend the standard conception of TDA on point clouds of unlabeled data, and develop new techniques to apply TDA to decision boundaries of labeled data.

The prior works we are aware of on TDA of decision boundaries include our own earlier work (Varshney & Ramamurthy, 2015) and Chen et al. (2019). In our prior work (Varshney & Ramamurthy, 2015), we use persistent homology to tune hyperparameters of radial basis function kernels and polynomial kernels. The contributions herein have greater breadth and theoretical depth as we detail below. Chen et al. (2019) consider the $0-$order homology of level sets of decision boundary function and use it to regularize classifier training. They do not consider higher order homology groups. A recent preprint also examines TDA of labeled data (Guss & Salakhutdinov, 2018), but approaches the problem as standard TDA on separate classes rather than trying to characterize the topology of the decision boundary. In Section 2 of the supplementary material (SM), we discuss how this approach can be fooled by the internal structure of the classes. There has also been theoretical work using rank of homology groups, known as Betti numbers, to upper and lower bound the number of layers and units of a neural network needed for representing a function (Bianchini & Scarselli, 2014). That work does not deal with data, as we do here. Moreover, its bounds are quite loose and not really usable in practice, similar in their looseness to the bounds for algebraic varieties (Basu et al., 2005; Milnor, 1964) cited in our prior work (Varshney & Ramamurthy, 2015) for polynomial kernel machines.

The main steps in a persistent homology analysis are as follows. We treat each data point as a node in a graph, drawing edges between nearby nodes — where nearby is according to a scale parameter. We form complexes from the simplices formed by the nodes and edges, and examine the topology of the complexes as a function of the scale parameter. The

[1] IBM Research, Yorktown Heights, NY, USA [2] Courant Institute, New York University, New York City, NY, USA. Correspondence to: Karthikeyan Natesan Ramamurthy <knatesa@us.ibm.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

topological features such as connected components, and holes of various dimensions that persist across scales are the ones that capture the underlying shape of the dataset. In all existing approaches to persistent homology, the scale parameter is a single global value that does not factor in the local scaling of the dataset, making the inference of Betti numbers from persistence brittle and difficult to automate.

Our main contributions are as follows:

1. We introduce a new simplicial complex construction called the labeled Čech complex that captures decision boundary topology. We provide theoretical conditions on the decision boundary and the data samples near the boundary that lead to the successful recovery of the homology of the decision boundary.

2. We propose a computationally efficient construction of decision boundary surfaces: the labeled Vietoris-Rips complex. We illustrate the need for local scaling to handle non-uniform sampling of data near the decision boundary and address this need by proposing a simplicial complex construction based on estimates of local scale using a k-nearest neighbors method.

3. We evaluate the merits of the above approaches using synthetic and real-world data experiments. Using synthetic data experiments, we show that the proposed approaches recover the homology even when there is extreme local scaling. Using the real-world application domains MNIST, FashionMNIST and CIFAR10, we show how these approaches can be used to evaluate the Decision Boundary Topological Complexity (DBTC) of deep neural network classifiers. One of the key advantages of our approach is that it provides a principled method to treat models and datasets in the same footing by generating *directly comparable* DBTC measures for both of them. These DBTC measures can then be used for selecting an appropriate pre-trained model suited to a novel dataset, without running the dataset through any of the models. This is quite useful in model market places (Bridgwater, 2018), where there can be privacy and security concerns that preclude running customer data on vendor models at the outset (See also Section 4). Our main finding in terms of model selection can be summarized as follows: *when choosing a pre-trained classifier model for a novel dataset, one whose DBTC matches that of the dataset yields good generalization.*

We defer detailed background on persistent homology and simplicial constructions for *unlabeled* point cloud data to Section 3 (SM). Throughout this work we assume the labels to be binary for simplicity; multi-class extensions can consider decision boundaries in one-vs-one, one-vs-all and Venn diagram constructions (Varshney & Willsky, 2010). In our experiments, we consider homology groups of dimensions 0 and 1 alone for computational efficiency, but the theory and methods are general to any dimension.

2. Labeled Čech Complex and Recovery Guarantees

In this section, we introduce the labeled Čech (LČ) complex and prove results on its use for recovering the homology of a decision boundary. The high-level idea is as follows: to recover the homology of a decision boundary, we must cover it such that the cover is its deformation retract. The practically- and computationally-oriented reader may safely proceed to Section 3 after noting the definition of the decision boundary and the proposed (computationally intractable) LČ complex.

2.1. Decision Boundary Manifold

Decision boundaries are collections of potentially multiple surfaces of dimension less than d in ambient spaces of dimension d, that divide a space into two classes. We define the overall probability space $\mathcal{Z}$ with the measure given by μ_z and the pdf p_Z. We assume two classes that can be conditioned from this space using the selector C; the pdfs being $p_X = p_{Z|C}(z|1)$ and $p_Y = p_{Z|C}(z|0)$. We denote the mixture probabilities as $p_C(0) = q$ and $p_C(1) = 1 - q$, such that $p_Z(z) = p_{Z|C}(z|1)p_C(1) + p_{Z|C}(z|0)p_C(0)$. By the Neyman-Pearson rule, the decision boundary manifold is defined by $\mathcal{M} = \{z \in \mathcal{Z} \mid p_Y = p_X\}$.

Let us define the extent of the distribution where the two classes are mixed by the set

$$\mathcal{D} = \{z \in \mathcal{Z} | p_{Z|C}(z|0) > 0, p_{Z|C}(z|1) > 0\}. \quad (1)$$

This is the set where both distributions have some mass.

2.2. Labeled Čech Complex

The homology of a manifold can be recovered by an appropriate random sampling and computing a Čech complex on the random samples. The same idea can be extended to the case of a decision boundary, which is a manifold at the intersection of the two classes. We need a construction which is homotopy equivalent to this manifold. To this end, we introduce the *labeled Čech complex.*

Definition 1. An (ϵ, γ)-labeled Čech complex, is a simplicial complex with a collection of simplices such that each simplex σ is formed on the points in the set S aided by the reference set W, when the following conditions are satisfied:

1. $\bigcap_{s_i \in \sigma} B_\epsilon(s_i) \neq \emptyset$, where s_i are the vertices of σ.

2. $\forall s_i \in \sigma, \quad \exists w \in W$ such that, $\|s_i - w\| \leq \gamma$.

Here, $B_\epsilon(s_i)$ denotes a ball of radius ϵ around the point s_i. This definition matches the usual Čech complex, but introduces the additional constraint that a simplex is induced only if all its vertices are close to some point in the reference set W. The second condition also implies that W is γ-dense in the vertices of the simplices of the (ϵ, γ)-LČ complex. Note that, if a set A is $\gamma-$dense in B, it means that for every $b \in B$, there exists $a \in A$, such that $\|b - a\| < \gamma$.

2.3. Recovery Guarantees

Now, we derive sufficient sampling conditions so that the LČ complex is homotopy equivalent to the decision boundary manifold and hence recovers it homology. The general idea is that when sufficient samples are drawn near $\mathcal{M}$, we can cover $\mathcal{M}$ using balls of radius r, and U deformation retracts to $\mathcal{M}$. The nerve of the covering will be homotopy equivalent to $\mathcal{M}$ according to the Nerve Lemma (Borsuk, 1948). The intuition is that when we have dense enough sampling, the nerve of the Čech complex is homotopy equivalent to the manifold (Niyogi et al., 2008). If the sampling is not sufficiently dense, we run into the danger of breaching the 'tubular neighborhood' of the manifold since the ϵ in the Čech complex has to be large. In our LČ complex, points from one class will be used to construct the actual complex, and the points from the other class will be used as the reference set W, based on Definition 1.

Sketch of the theory: Lemma 1 shows the equivalence of the LČ complex to a particular Čech complex on unlabeled data, helping us build our theory from existing results in (Niyogi et al., 2008). Theorem 1 lower bounds the sample size needed to cover two sets of sets, laying the ground for our main sample complexity result. Theorem 2 provides the sample complexity for a dense sampling of the decision boundary manifold, and the main result in Theorem 3 gives the sufficient conditions under which an LČ complex on the sampled points from the two classes will be homotopy equivalent to the decision boundary.

Let us assume that:

- The decision boundary is a manifold $\mathcal{M}$ with condition number $1/\tau$. This means that the open normal bundle about $\mathcal{M}$ of radius τ is embedded in $\mathbb{R}^d$. In other words, the normal bundle is non self-intersecting. In more practical terms, a large τ corresponds to a well-conditioned manifold with low curvature (do Carmo, 1992).
- $\mathcal{D}$ is contained in the tubular neighborhood of radius r around $\mathcal{M}$, i.e., $\mathcal{D} \subset \text{Tub}_r(\mathcal{M})$.
- For every $0 < s < r$, the mass around a point p in $\mathcal{M}$ is at least $k_s^{(c)}$ in both classes. There is sufficient mass in both classes:

$$\inf_{p \in \mathcal{M}} \mu_c(B_\epsilon(p)) > k_s^{(c)} \quad \forall c \in \{0, 1\}. \tag{2}$$

Lemma 1. *As ϵ varies from 0 to ∞, a filtration is induced on the (ϵ, γ)-LČ complex for a fixed γ.*

Proof. Fixing γ, we choose $S_\gamma \subseteq S$, such that W is γ-dense in S_γ (See Defn. 1 for definitions of S and W, and also a note on $\gamma-$density). Therefore, the (ϵ, γ)-LČ complex on S is equivalent to an ϵ-Čech complex on S_γ, and hence varying ϵ induces a filtration. □

Remark. The (ϵ, γ)-Čech complex can be used to delineate the decision boundary by choosing S to be the samples of one class and W to be the other class.

Given sufficient samples in S and W, a union of ϵ-balls on S_γ will be homotopy equivalent to $\mathcal{M}$, when ϵ is chosen such that there is a 'good covering'. Since homotopy implies same homology, this is how we use the LČ complex to identify the homology of the decision boundary.

Theorem 1. *Let $\{A_i\}_{i=1}^{l_a}$ and $\{B_j\}_{j=1}^{l_b}$ be two sets of measurable sets. Let μ_x and μ_y be the probability measures on $\bigcup_{i=1}^{l_a} A_i$ and $\bigcup_{j=1}^{l_b} B_j$, respectively, such that $\mu_x(A_i) > \alpha_x, \forall i \in \{1, 2, \dots, l_a\}$ and $\mu_y(B_i) > \alpha_y, \forall j \in \{1, 2, \dots, l_b\}$. Let μ_x and μ_y be the component measures of μ_z, such that $\mu_z(F) = q\mu_x(F) + (1-q)\mu_y(F)$, q and $1-q$ being the mixture probabilities. Let $\overline{z} = \{z_1, z_2, \dots, z_n\}$ be the set of n i.i.d. draws according to μ_z, which can be partitioned into two sets $\overline{x}$ and $\overline{y}$ which contain the samples from the measures μ_x and μ_y. Then, if*

$$n \geq \max\left(\frac{1}{\alpha_x q}\left(\log 2l_a + \log\frac{1}{\delta}\right), \frac{1}{\alpha_y(1-q)}\left(\log 2l_b + \log\frac{1}{\delta}\right)\right) \tag{3}$$

we are guaranteed with probability greater than $1 - \delta$ that

$$\forall i, \overline{x} \cap A_i \neq \emptyset \quad \textit{and} \quad \forall j, \overline{y} \cap B_j \neq \emptyset. \tag{4}$$

See Section 1 (SM) for proof.

Lemma 2. *For three sets S, W, and U, if S is r-dense in U and W is t-dense in U, there exists an $\hat{S} \subseteq S$, such that the following hold:*

1. *$\hat{S}$ is r-dense in U,*
2. *U is r-dense in $\hat{S}$,*
3. *W is $(r + t)$-dense in $\hat{S}$.*

Proof. If S is r-dense in U, for every $u \in U$, there exists an $s \in S$ such that $\|u - s\| < r$. Now, let $\hat{S} \subseteq S$, $\hat{S} =$

$\{s \in S \mid \|u-s\| < r, u \in U\}$, i.e., for each element $\hat{s} \in \hat{S}$, we have at least one $u \in U$ such that $\|u - \hat{s}\| < r$ and vice-versa. This proves item 1 and item 2. Since for each u, we have at least one $w \in W$, such that $\|u-w\| < t$. Hence, by the triangle inequality, for each $\hat{s} \in \hat{S}$, we have at least one $w \in W$ such that $\|s - w\| < (r+t)$. □

Theorem 2. *Let $N_{r/2}$ and $N_{s/2}$ be the $r/2$ and $s/2$ covering numbers of the manifold $\mathcal{M}$. Let G and H be two sets of points in $\mathcal{M}$ of sizes $N_{r/2}$ and $N_{s/2}$ such that $B_{r/2}(g_i), g_i \in G$, and $B_{s/2}(h_j), h_j \in H$ are the $r/2$- and $s/2$-covers. Let $\overline{z}$ be generated by i.i.d. sampling from μ_z whose two component measures satisfy the regularity properties in (2), and have mixing probabilities q and $1-q$ for $q > 0$. Let the two component samples be $\overline{x}$ and $\overline{y}$. Then if*

$$|\overline{z}| > \max\Big(\frac{1}{qk^{(0)}_{r/2}}\left(\log\left(2N_{r/2}\right)+\log\left(\tfrac{1}{\delta}\right)\right), \frac{1}{(1-q)k^{(0)}_{s/2}}\left(\log\left(2N_{s/2}\right)+\log\left(\tfrac{1}{\delta}\right)\right)\Big),$$

with probability greater than $1-\delta$, $\overline{x}$ will be r-dense in $\mathcal{M}$, and $\overline{y}$ will be s-dense in $\mathcal{M}$.

Proof. Letting $A_i = B_{r/2}(g_i)$, and $B_j = B_{s/2}(h_j)$, apply the previous Theorem. Hence, with probability greater than $1-\delta$, each of A_i and B_j are occupied by at least one of $x_i \in \overline{x}$, and $y_j \in \overline{y}$ respectively. There it follows that for any $p \in \mathcal{M}$, there is at least one $x \in \overline{x}$ and $y \in \overline{y}$ such that $\|p-x\| < r$, and $\|p-y\| < s$. Thus, with high probability, $\overline{x}$ is r-dense in $\mathcal{M}$ and $\overline{y}$ is s-dense in $\mathcal{M}$. □

Now we extend Theorem 7.1 in Niyogi et al. (2008) to the case of the LČ complex and provide the main conditions under which the homology of the decision boundary can be recovered.

Theorem 3. *Let $N_{r/2}$ and $N_{s/2}$ be the $r/2$ and $s/2$ covering numbers of the submanifold $\mathcal{M}$ of $\mathbb{R}^N$. Let $\overline{z}$ be generated by i.i.d. sampling from μ_z whose two component measures satisfy the regularity properties in (2), and have mixing probabilities q and $1-q$ for $q > 0$. Let the two component samples be $\overline{x}$ and $\overline{y}$. Then if*

$$|\overline{z}| > \max\Big(\frac{1}{qk^{(0)}_{r/2}}\left(\log\left(2N_{r/2}\right)+\log\left(\tfrac{1}{\delta}\right)\right), \frac{1}{(1-q)k^{(0)}_{s/2}}\left(\log\left(2N_{s/2}\right)+\log\left(\tfrac{1}{\delta}\right)\right)\Big),$$

with probability greater than $1-\delta$, the $(\epsilon, r+s)$-LČ complex will be homotopy equivalent to $\mathcal{M}$, if: (a) $r < (\sqrt{9}-\sqrt{8})\tau$, and (b) $\epsilon \in \left(\frac{(r+\tau)-\sqrt{r^2+\tau^2-6\tau r}}{2}, \frac{(r+\tau)+\sqrt{r^2+\tau^2-6\tau r}}{2}\right)$.

Proof. From Lemma 2, we know that when $\overline{x}$ is r-dense in $\mathcal{M}$, and $\overline{y}$ is s-dense in $\mathcal{M}$, we have $\tilde{x} \subseteq \overline{x}$ which is also r-dense in $\mathcal{M}$ and $\overline{y}$ is $(r+s)$-dense in $\tilde{x}$. Also, from Lemma 1, the $(\epsilon, r+s)$-LČ complex on $\overline{x}$ with the reference set $\overline{y}$ is equivalent to the ϵ-Čech complex on $\tilde{x}$.

Since $\tilde{x}$ is r-dense on $\mathcal{M}$, it follows from Theorem 7.1 in (Niyogi et al., 2008) that this ϵ-Čech on $\tilde{x}$ will be homotopy equivalent to $\mathcal{M}$ if the conditions on r and ϵ are satisfied. □

3. Labeled Vietoris-Rips Complexes

In this section, we propose two computationally-tractable constructions for simplicial complexes of the decision boundary: one we name the plain labeled Vietoris-Rips complex and the other we name the locally scaled labeled Vietoris-Rips complex. We illustrate the need for the locally scaled version.

3.1. Notation

Let us start with a labeled discrete sample $\{(z_1, c_1), \ldots, (z_n, c_n)\}$ where $z_i \in \mathbb{R}^d$ is the data point and $c_i \in \{0, 1\}$ is its class label. Given a data point z_i, we define its neighborhood as the set $\mathcal{N}_\theta(z_i)$ where θ is a scalar neighborhood parameter. The neighbors are restricted to data points whose class c_j is not the same as c_i. Our neighborhood construction is symmetric by definition, hence $z_j \in \mathcal{N}_\theta(z_i) \Leftrightarrow z_i \in \mathcal{N}_\theta(z_j)$. This results in a bipartite graph G_θ.

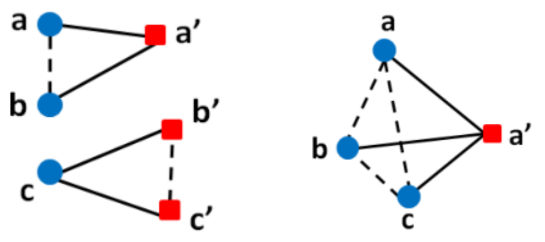
We use Betti numbers to describe the topology of the decision boundary. The i^{th} Betti number β_i is the rank of the homology group H_i of dimension i and is a count of connected components, holes, or cavities of dimension i.

3.2. Two Complexes

To induce a simplicial complex with simplices of order greater than one from the bipartite graph G, we connect all 2-hop neighbors.[1] Since the original edges are only between points in opposing classes, all 2-hop neighbors belong to the same class. Consider the examples in Figure 1. In the first example, we start with three points in a two-dimensional space where all points are within ϵ of each other. Two share a class label and are thus not initially connected by an edge. The initial graph $\mathbf{A}$ has two line segment simplices. After including the graph walk, an intraclass edge is introduced. Now $\tilde{\mathbf{A}}$ has a triangle simplex. The second example is similar, but has four points in three-dimensional space, with three of the four points sharing a class label. Here we form a tetrahedron after introducing the length two graph walk edges.

This new graph is defined to be one-skeleton of the decision

[1] We find two hops to be sufficient. Larger hops may be considered, but could introduce spurious cycles.



(a) (b)

Figure 1. (a) A simplicial complex with two 2-simplices from a bipartite graph between circle and square classes generated using length-2 walks (dotted lines), (b) a complex created with one 3-simplex using the same approach.

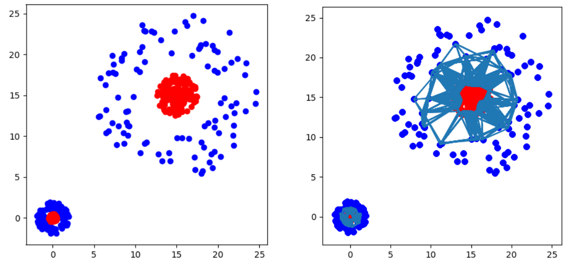
boundary complex. We create a simplicial complex from this one-skeleton using the standard Vietoris-Rips induction (Zomorodian, 2010): a simplex of dimension $r+1$ is inductively included in the complex if all its r-dimensional faces are included. We call this the labeled Vietoris-Rips (LVR) complex $\mathcal{V}_\theta$.

Our construction is such that, by definition, for $\theta_2 \geq \theta_1$, there is an inclusion $G_{\theta_2} \supseteq G_{\theta_1}$. Given this inclusion relationship in the bipartite graphs, we obtain a filtration as we vary θ, i.e., for $\theta_2 \geq \theta_1$, $\mathcal{V}_{\theta_2} \supseteq \mathcal{V}_{\theta_1}$. We provide two approaches for creating the LVR complex and its filtration.

Plain LVR (P-LVR) Complex: We set θ to be the radius parameter ϵ and define $\mathcal{N}_\theta(z_i)$ as the set of points $\{z_j\}_{c_j \neq c_i, \|z_i - z_j\| \leq \theta}$. Persistent homology is obtained by varying the radius parameter ϵ.

Locally Scaled LVR (LS-LVR) Complex: We set θ to be κ, the multiplier to the local scale and define $\mathcal{N}_\theta(z_i)$ as the set of points $\{z_j\}_{c_j \neq c_i, \|z_i - z_j\| \leq \kappa\sqrt{\rho_i \rho_j}}$, where ρ_i is the local scale of z_i. This is defined to be the radius of the smallest sphere centered at z_i that encloses at least k points from the opposite class. In this complex, persistent homology is obtained by varying the local scale multiplier, κ. LS-LVR construction is based on the generalization of CkNN graph introduced in (Berry & Sauer, 2019) to labeled data. Similar *weighted simplicial constructions* where the radius of the ball around a point depends on the point itself have been discussed in (Bell et al., 2017), where the authors show stability with respect to persistence diagrams (PDs) for small perturbations of the point cloud. Local scaling helps in making the persistent homology computation invariant to the sampling density, as illustrated in the next section.

After the LVR filtrations have been obtained, persistent homology of the decision boundaries can be estimated using standard approaches (Edelsbrunner & Harer, 2008; Zomorodian & Carlsson, 2005), and represented using barcodes or persistence diagrams (Edelsbrunner & Morozov, 2012).



(a) (b)

Figure 2. (a) A 2-class dataset with *red* and *blue* classes, and (b) the LS-LVR decision boundary complex at $\kappa = 1.005$.

3.3. Illustration of Homology Group Recovery

We illustrate these two approaches for constructing decision boundary complexes and estimating their persistent homology using a two-dimensional, two-class dataset given in Figure 2(a). The two decision boundaries are homotopy equivalent to two circles that separate the classes, and hence the true Betti numbers of the decision boundaries for this data are: $\beta_0 = 2$, and $\beta_1 = 2$. The sampling is non-uniform, with the smaller disk and annulus having more density than the larger ones.

We compute the persistent homology using the P-LVR and LS-LVR complexes. With P-LVR, we vary the radius parameter ϵ from 0 to 10, and with LS-LVR, we vary the local scale multiplier κ from 0.5 to 1.5. The local scale ρ is computed with $k = 5$ neighbors. Figure 2(b) shows a LS-LVR complex at scale 1.005 that accurately recovers the Betti numbers of the decision boundary. Note that the varying sampling densities of the two regions in the decision boundary (small and big circles) get normalized due to local scaling.

Figure 3 shows the PDs as well as the Betti numbers for different scales using the two complexes.[2] The LS-LVR construction recovers both β_0 and β_1 accurately for κ slightly greater than 1 and persists until κ is slightly less than 1.2. Around this value, one of the holes closes and a little later the other hole collapses as well. The resulting two simply connected components persist until $\kappa = 1.5$.

In contrast, for the P-LVR complex, the H_0 and H_1 groups first come to life at $\epsilon = 0.9$ for the smaller decision bound-

[2] Note that the PD for H_0 groups shows all the groups, whereas the Betti numbers in Figures 3(b) and 3(d) only show the number of non-trivial homology groups. Non-trivial H_0 groups are defined to be those that have more than one data point i.e., the number of simply connected components with size more than 1. Including trivial homology groups is meaningless when computing the topology of decision boundaries since decision boundaries are defined only across classes.

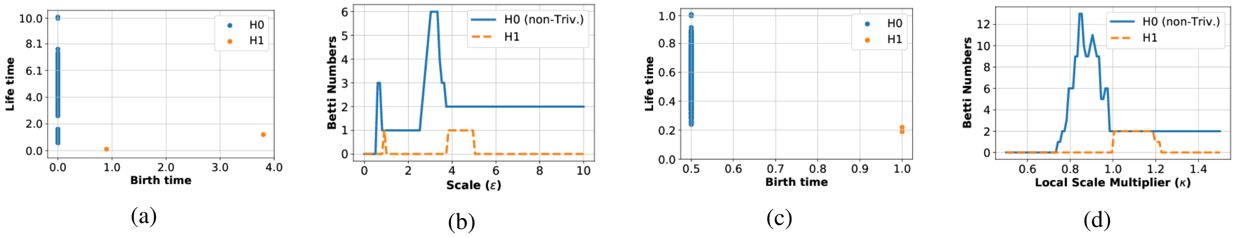


Figure 3. For the data in Figure 2. (a) Persistence diagram and (b) Betti numbers as a function of scale using P-LVR, and (c) persistence diagram and (d) Betti numbers using LS-LVR. The axes of the persistence diagrams are *birth time* and *life time = death time-birth time*.

ary component. The H_1 group vanishes almost immediately. At $\epsilon = 0.38$, the H_0 and H_1 groups for the larger decision boundary component come to life, persisting for 0.12. The overall topology ($\beta_0 = 2, \beta_1 = 2$) is not captured at any one scale due to varying sizes of homological features as well as non-uniform sampling. The widely varying life times for homological features make it hard to choose a threshold for estimating the correct number of homology groups. This is not a problem with LS-LVR since the H_1 groups appear clustered together in the PD. Another benefit of LS-LVR is that non-noisy homology groups appear around $\kappa = 1$, the natural local scale of the data. This does not hold true for the P-LVR complex. The actual complexes for various scales with the two constructions are given in Section 4 (SM).

4. Experiments

We perform experiments with synthetic and high-dimensional real-world datasets to demonstrate: (a) the effectiveness of our approach in recovering homology groups accurately, and (b) the utility of this method in discovering the Decision Boundary Topological Complexity (DBTC) of neural networks and their potential use in choosing pre-trained models for a new dataset.

Marketplace for pre-trained models is an upcoming trend in the machine learning/AI industry (Bridgwater, 2018). In this setting, pre-trained models are available from vendors, and the customers are expected to choose an appropriate model for their dataset from this model marketplace. The data may be sensitive, and the models may have proprietary technology, so it may not be possible for customers to run their datasets on vendor models. Our model selection application is a natural fit here, since we will extract DBTC measures for customer data and vendor models and match them up to find the best model for the data.

4.1. Implementation Notes

In all experiments, to limit the number of simplices, we upper bound the number of neighbors used to compute the neighborhood graph to 20. We adopt several approaches to make our implementations efficient. We describe them briefly:

- We use ϵ-neighborhood graphs to compute the LVR complexes, but to limit the number of simplices, we restrict the number of nearest neighbors for any point to 20. We then symmetrize the graph and use it for obtaining the P-LVR and LS-LVR constructions.
- The neighborhood graphs are computed efficiently using Cython code interfaced to the main Python package that we developed.
- We estimate the distance matrices for the LVR constructions and use the efficient Ripser package (Bauer, 2016) and its Python interface (Nathaniel Saul, 2019) to obtain the persistence diagrams.
- The entire pipeline (neighborhood graph construction and LVR estimation) to estimate the Betti numbers β_0 and β_1 for two classes runs in less than 1 minute for about 1000 points per class (the standard size of our test datasets). The program runs in a single core using less than 500MB of RAM in a standard computer.

Implementations of the approaches proposed in this work are available at: `https://github.com/nrkarthikeyan/topology-decision-boundaries`

4.2. Synthetic Data: Homology Group Recovery

The first experiment demonstrates the effectiveness of our approach in recovering homology groups of complex synthetic data with wide variations in sizes of topological features (Figure 4). The decision boundary is homotopy equivalent to 25 circles ($\beta_0 = 25, \beta_1 = 25$). From Figures 5(c) and 5(d), it is clear that the LS-LVR complex shows similar persistence for all the 25 H_1 groups irrespective of their varying sizes in the dataset. Observe the clumping in the PD, and the presence of a lone noisy H_1 group with almost zero life time. The P-LVR complex also recovers the 25 H_1 groups, but does so at different times (Figures 5(a) and 5(b)). From the PD, we can see that there are five rough clumps

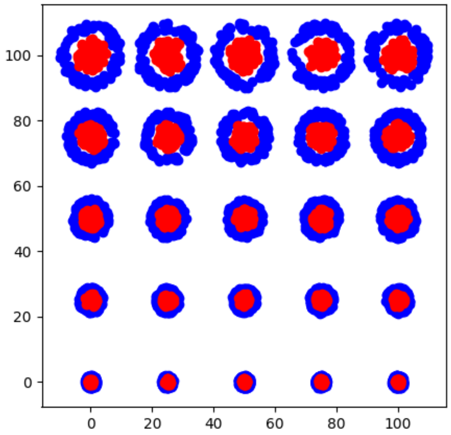


Figure 4. A 2-class dataset with $\beta_0 = 25, \beta_1 = 25$. Notice the wide variation in sizes of topological features.

of H_1 groups, around birth times $\{1, 2, 3, 4, 5\}$, each containing five H_1 groups. The birth times correspond to the radii of the five groups of decision boundaries in Figure 4. The staggered recovery of topology with the P-LVR complex makes it hard to fix a noise threshold on life times to estimate the correct Betti numbers.

4.3. Real-World Data: Complexity Estimation and Model Selection

We demonstrate how topological complexity can be used to guide selecting appropriate pre-trained models for a new dataset. We use only LS-LVR complexes for estimating topological complexities. We consider three application domains for our evaluation: MNIST (Lecun & Cortes, 2009), FashionMNIST (Xiao et al., 2017) and CIFAR10 (Krizhevsky & Hinton, 2009). All three applications have 10 classes and $50,000$ training and $10,000$ test images. Each instance of MNIST and FashionMNIST is a 28×28 grayscale image, whereas each instance of CIFAR10 is a 32×32 color image. We construct $\binom{10}{2} = 45$ binary classification datasets from each application domain, one for each combination of two classes. We then train individual binary classifiers for these 45 datasets per application using the standard CNN architecture provided in `https://github.com/pytorch/examples/tree/master/mnist` for MNIST and FashionMNIST, and the VGG CNN - configuration D for CIFAR10 (Simonyan & Zisserman, 2014).

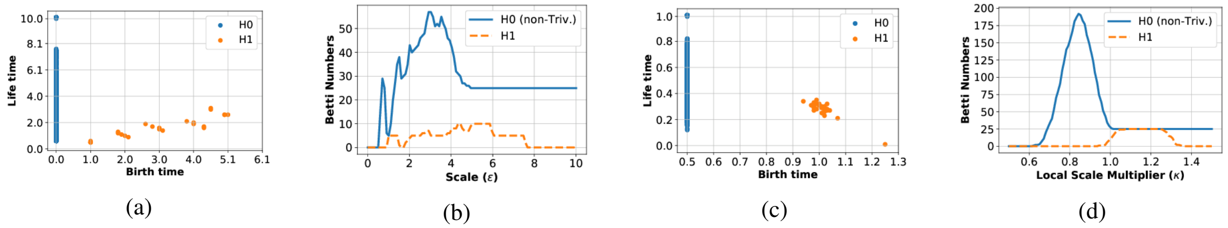
4.3.1. Estimating the DBTC of Data and Models

The DBTC measures that are described here provide a way to treat data and models in an equal footing hence providing a way to compare them easily. The DBTC of a pre-trained model $f_i(\cdot)$ can be obtained using the test data features $(z_{i,.})$ and the labels predicted by the model ($(f(z_{i,.}))$. This labeled dataset, which is a sample-level representation of the model, is given as $Z_i^{(m)} = \{(z_{i,1}, f_i(z_{i,1})), \ldots, (z_{i,n_i}, f_i(z_{i,n_i}))\}$ and it characterizes the decision boundary. For a novel dataset, given by $Z_i^{(d)} = \{(z_{i,1}, c_{i,1}), \ldots, (z_{i,n_i}, c_{i,n_i})\}$, the DBTC can be directly estimated using the input features $(z_{i,.})$ and the true labels $(c_{i,.})$. Now, we can compare the DBTCs of the novel dataset to that of the models to evaluate the suitability of the pre-trained model to the novel dataset. For some dataset or model indexed by i, let us denote the Betti numbers for H_0 and H_1 at scale κ as $\beta_{0,\kappa}(i)$, and $\beta_{1,\kappa}(i)$ respectively. We will also denote the persistence diagrams for H_0 and H_1 as $P_0(i)$ and $P_1(i)$.

We provide five different measures for quantifying the DBTC of the decision boundaries in $Z_i^{(m)}$ or $Z_i^{(d)}$: (a) The total lifetime of H_0 groups given by $\sum_\kappa \beta_{0,\kappa}$, (b) the total lifetime of H_1 groups given by $\sum_\kappa \beta_{1,\kappa}$, (c) the D_p divergence (Berisha et al., 2016) between the two classes, (d) the H_0 persistence, P_0 and, (e) the H_1 persistence, P_1. The homology-based measures (a, b, d, e) are obtained from the LS-LVR complexes, and the D_p divergence between the two component class distributions can be efficiently estimated by constructing a minimum spanning tree on $Z_i^{(m)}$ or $Z_i^{(d)}$ (Berisha et al., 2016). The first two measures $\sum_\kappa \beta_{0,\kappa}$ and $\sum_\kappa \beta_{1,\kappa}$ are also referred to as *degree-0 total persistence* in Cohen-Steiner et al. (2010), and the persistence diagrams (d and e) are well-known measures to quantify homology. Since the measures (a), (b), (c) are scalars, for two decision boundaries, they can be compared by computing the absolute of differences. For the persistence diagrams, comparison is performed using sliced Wasserstein distance discussed in Carriere et al. (2017).

4.3.2. Matching Novel Datasets to Pre-Trained Models

Let us start with an example of how we can use these measures to match novel datasets to pre-trained models. Our novel dataset is MNIST handwritten digit 0 vs. handwritten digit 4, whose DBTC measure (b) - lifetime of H_1 groups - is 91. Then we look for pre-trained model complexities that are similar. Not surprisingly, the closest is the pre-trained model 0 vs. 4, which has a model complexity 91. The 0 vs. 9 pre-trained model has a similar complexity of 112. If we select the 0 vs. 4 model, we achieve 99.95% accuracy on 0 vs. 4 data, and if we select the 0 vs. 9 model, we also achieve a high accuracy of 96.08%. If we select a model that is not well-matched to the data complexity, for example the 0 vs. 5 model with complexity 300, we achieve a low accuracy on 0 vs. 4 data of 63.41%. The DBTC measures (a) and (b) for datasets and models, and accuracies are listed in Section 5 (SM) and Section 6 (SM), respectively.



(a)

(b)

(c)

(d)

Figure 5. For the data in Figure 4: (a) Persistence diagram and (b) Betti numbers as a function of scale using P-LVR, and (c) persistence diagram and (d) Betti numbers using LS-LVR. The axes of the persistence diagrams are *birth time* and *life time* = *death time-birth time*.

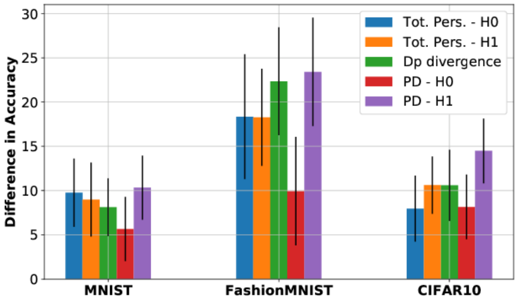


Figure 6. Accuracy improvement or reduction in choosing pre-trained classifiers with topological complexity close to the dataset versus complexity far from the dataset. Complexity measures used: (a) Total persistence for H_0 groups, (b) Total persistence for H_1 groups, (c) D_p divergence, (d) Persistence diagrams of H_0 groups, (e) Persistence diagrams of H_1 group. The black lines show the 95% confidence interval.

Now let us conduct an experiment to see whether the example above holds in general. Treating each of the 45 datasets as the novel dataset, we select 5 pre-trained models that are the closest and 5 models that are the farthest in topological complexity. We evaluate these classifiers on the novel dataset and obtain the average difference in classification accuracy between the closest and farthest classifiers. If the difference in accuracy is significantly greater than zero, it means that using classifiers that have similar DBTC as the dataset is beneficial. If the difference in accuracy is close to zero, it shows that there is no benefit in using DBTC to guide the choice of the classifier. If it is significantly less than zero, it means that classifiers which do not have similar DBTC are better suited for the novel dataset.

Armed with this intuition, we can interpret Figure 6. The bars show the average accuracy difference obtained by repeating the above experiment on the 45 two-class datasets in each of CIFAR10, MNIST and FashionMNIST. The black lines show the 95% confidence interval using a one-sample t-test. If the black line is completely above (below) 0, with a p-value less than 0.05, the null hypothesis that the accuracy difference is less than or equal to (greater than or equal to) 0 can be rejected. If the black line intersects 0, we cannot reject the null hypothesis that the accuracy difference is 0, at a significance level of 0.05. From the bars, we see that pre-trained classifier models that have similar DBTC as the novel dataset show higher performance in that dataset on the novel dataset for all three complexity measures. This confirms our main claim that choosing classifier models that have similar DBTC as the novel dataset is beneficial. Furthermore, the DBTC measure (e) - PD for H_1 groups - is the best for model selection task. The baseline D_p divergence measure also shows a strong performance.

5. Conclusion

In this paper, we have investigated the use of topological data analysis in the study of labeled point clouds of data encountered in supervised classification. In contrast to Guss & Salakhutdinov (2018), which simply applies known, standard, persistent homology inference methods to different classes of data separately and does not scale to high dimensions, we introduce new techniques and constructions for characterizing *decision boundaries* and apply them to several commonly used datasets in deep learning. We propose and theoretically analyze the *labeled* Čech complex, deriving conditions on recovering the decision boundary's homology with high probability based on the number of samples and the condition number of the decision boundary manifold.

Furthermore, we have proposed the computationally-tractable *labeled* Vietoris-Rips complex and extended it to account for variation in the local scaling of data across a feature space. We have used this complex to provide a complexity quantification of pre-trained models and datasets that is able to correctly identify the complexity level below which a pre-trained model will suffer in its ability to generalize to a given dataset. This use has increasing relevance as model marketplaces become the norm.

Acknowledgements

The authors thank the anonymous reviewers for their helpful comments. The authors also thank Visar Berisha for discussions regarding experiments using the D_p divergence measure. Figure 1 (©2015 IEEE) is reprinted with permission, from Varshney & Ramamurthy (2015).

References

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al. A closer look at memorization in deep networks. In *Proc. Int. Conf. Mach. Learn.*, pp. 233–242, 2017.

Basu, S., Pollack, R., and Roy, M.-F. Betti number bounds, applications and algorithms. *Current trends in combinatorial and computational geometry: papers from the special program at MSRI*, 52:87–97, 2005.

Bauer, U. Ripser: Efficient computation of Vietoris-Rips persistence barcodes, 2016. URL https://github.com/Ripser/ripser.

Bell, G., Lawson, A., Martin, J., Rudzinski, J., and Smyth, C. Weighted persistent homology. arXiv:1709.00097, 2017.

Berisha, V., Wisler, A., Hero III, A. O., and Spanias, A. Empirically estimable classification bounds based on a nonparametric divergence measure. *IEEE Trans. Signal Processing*, 64(3):580–591, 2016.

Berry, T. and Sauer, T. Consistent manifold representation for topological data analysis. *Found. Data Sci.*, 1(1):1–38, 2019.

Bianchini, M. and Scarselli, F. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE Trans. Neural Netw. Learn. Syst.*, 25(8):1553–1565, August 2014.

Borsuk, K. On the imbedding of systems of compacta in simplicial complexes. *Fundamenta Mathematicae*, 35(1): 217–234, 1948.

Bridgwater, A. Enough Training, Let's Get Down To The AI Supermarket. *Forbes*, Sep 2018. URL https://www.forbes.com/sites/adrianbridgwater/2018/09/18/enough-training-lets-get-down-to-the-ai-supermarket/#5f13cdbc10c3.

Carlsson, G. Topology and data. *Bull. Amer. Math. Soc.*, 46 (2):255–308, April 2009.

Carriere, M., Cuturi, M., and Oudot, S. Sliced Wasserstein kernel for persistence diagrams. In *Proc. Int. Conf. Mach. Learn.*, pp. 664–673, 2017.

Chen, C., Ni, X., Bai, Q., and Wang, Y. A topological regularizer for classifiers via persistent homology. In *Proc. Int. Conf. Artif. Intell. Stat.*, pp. 2573–2582, 2019.

Cohen-Steiner, D., Edelsbrunner, H., Harer, J., and Mileyko, Y. Lipschitz functions have l_p-stable persistence. *Found. Comput. Math.*, 10(2):127–139, April 2010.

do Carmo, M. P. *Riemannian geometry, Translated by Francis Flaherty*. Birkhäuser, 1992.

Edelsbrunner, H. and Harer, J. Persistent homology—a survey. *Contemp. Math.*, 453:257–282, 2008.

Edelsbrunner, H. and Morozov, D. Persistent homology: theory and practice. Technical report, Lawrence Berkeley National Laboratory, Berkeley, CA, USA, 2012.

Guss, W. H. and Salakhutdinov, R. On characterizing the capacity of neural networks using algebraic topology. arXiv:1802.04443, 2018.

Ho, T. K. and Basu, M. Complexity measures of supervised classification problems. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(3):289–300, March 2002.

Kawaguchi, K., Kaelbling, L. P., and Bengio, Y. Generalization in deep learning. arXiv:1710.05468, 2017.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Lecun, Y. and Cortes, C. The MNIST database of handwritten digits, 2009. URL http://yann.lecun.com/exdb/mnist/.

Milnor, J. On the Betti numbers of real varieties. *Proc. Am. Math. Soc.*, 15(2):275–280, April 1964.

Nathaniel Saul, C. T. Scikit-TDA: Topological Data Analysis for Python, 2019. URL https://doi.org/10.5281/zenodo.2533369.

Niyogi, P., Smale, S., and Weinberger, S. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, 39(1-3):419–441, 2008.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

Varshney, K. R. and Ramamurthy, K. N. Persistent topology of decision boundaries. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 3931–3935, Brisbane, Australia, April 2015.

Varshney, K. R. and Willsky, A. S. Classification using geometric level sets. *J. Mach. Learn. Res.*, 11:491–516, February 2010.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747, 2017.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. arXiv:1611.03530, 2016.

Zomorodian, A. Fast construction of the Vietoris-Rips complex. *Comput. Graph.*, 34(3):263–271, 2010.

Zomorodian, A. and Carlsson, G. Computing persistent homology. *Discrete Comput. Geom.*, 33(2):249–274, February 2005.