

# A Topological Perspective on Neural Network Training

Thomas Walker

Summer 2023

## Contents

<b>1 Introduction</b>	<b>1</b>
<b>2 Persistent Homology</b>	<b>1</b>
<b>3 Binary Classification</b>	<b>3</b>
<b>4 Architectural Implications on Data Topology</b>	<b>3</b>

## 1 Introduction

In [1] deep neural networks are investigated by understanding how the layers manipulate the input's topology. Persistent homology is used to understand how topological features of the input data evolve through the layers of the network, and how architectural properties of the network contribute to this evolution. In [1] a framework for this investigation is developed and some preliminary conclusions are given. It does not attempt to explain network generalization from the topological lens.

## 2 Persistent Homology

**Definition 2.1.** A  $k$ -dimensional simplex  $\sigma$  in  $\mathbb{R}^d$  is the convex hull of  $k + 1$  affinely independent points  $v_0, \dots, v_k \in \mathbb{R}^d$ , denoted  $\sigma = [v_0, \dots, v_k]$ .

The faces of a  $k$ -simplex are simplices of dimension 0 to  $k - 1$  formed by convex hulls of proper subsets of the vertex set  $\{v_0, \dots, v_k\}$ .

**Definition 2.2.** An  $m$ -dimensional geometrical simplicial complex  $K$  in  $\mathbb{R}^d$  is a finite collection of simplices in  $\mathbb{R}^d$  of dimension at most  $m$  that are

1. Any intersection between two simplices in  $K$  is necessarily a face of both of them, and
2. includes all faces of all its simplices.

**Definition 2.3.** An abstract simplicial complex is a list of simplices  $K = \{\sigma_1, \dots, \sigma_n\}$  such that if  $\tau \subseteq \sigma \in K$  then  $\tau \in K$ .

**Definition 2.4.** Let  $C_0, \dots, C_d$  be vector spaces over  $\mathbb{F}_2$ . Let boundary operators  $\delta_k : C_k \rightarrow C_{k-1}$  be linear maps satisfying

$$\delta_k \circ \delta_{k-1} = 0$$

for all  $k = 1, \dots, d$ .

**Definition 2.5.** A chain complex is a sequence

$$0 \xrightarrow{\delta_{d+1}} C_d \dots C_1 \xrightarrow{\delta_0} 0,$$

where  $C_{d+1} = C_{-1} = 0$ . The elements in the image of  $\delta_k$  are called boundaries and elements of the kernel of  $\delta_{k-1}$  are called cycles,

$$B_k := \text{im}(\delta_{k+1}) \subseteq \ker(\delta_k) =: Z_k.$$

**Definition 2.6.** The  $k^{\text{th}}$  homology group is the quotient vector space

$$H_k := Z_k / B_k, \quad \text{for } k = 0, \dots, d.$$

The homology classes are the equivalence classes

$$[z] = z + B_k = \{z + b \in Z_k : b \in B_k\}.$$

**Definition 2.7.** The  $k^{\text{th}}$  Betti number of  $K$  is  $\beta_k := \dim(H_k)$ .

**Remark 2.8.** The  $k^{\text{th}}$  Betti number of  $K$  counts the number of  $k$ -dimensional holes in  $K$ .

Let  $K^{(k)} = \{\sigma_1, \dots, \sigma_m\}$  be the set of all  $k$ -dimensional simplices in  $K$ . When working over  $\mathbb{F}_2$ , let  $C_k(K)$  be the vector space with basis  $K^{(k)}$ . Define the boundary operator  $\delta_k : C_k(K) \rightarrow C_{k-1}(K)$  for a  $k$ -simplex  $\sigma = [v_0, \dots, v_k]$  as

$$\delta_k(\sigma) = \delta_k \left( \sum_{j=1}^m n_j \sigma_j \right) = \sum_{j=1}^m n_j \delta_k(\sigma_j).$$

Working over  $\mathbb{F}_2$  ensures that  $H_k(K) \cong \mathbb{F}_2^{\beta_k}$  and

$$\beta_k(K) = \dim(H_k(K)) = \text{null}(\delta_k) - \text{rank}(\delta_{k+1}) \quad k = 0, \dots, d.$$

Let  $m_k = |K^{(k)}|$ , then the number of simplices in a  $d$ -dimensional simplicial complex is bounded by

$$|K| \leq \sum_{i=1}^d \binom{m_0}{i+1}.$$

**Definition 2.9.** For  $K_1$  and  $K_2$  two abstract simplicial complexes a simplicial map  $f : K_1^{(0)} \rightarrow K_2^{(0)}$  is such that for  $\sigma = [v_0, \dots, v_k] \in K_1$  it follows that  $[f(v_0), \dots, f(v_k)] \in K_2$ . This induces a map between complexes

$$f : C_k(K_1) \rightarrow C_k(K_2), \quad \sum_{j=1}^m n_j \sigma_j \mapsto \sum_{j=1}^m n_j f(\sigma_j),$$

which induces a map between homologies

$$H_k(f) : H_k(K_1) \rightarrow H_k(K_2), \quad \left[ \sum_{j=1}^m n_j \sigma_j \right] \mapsto [n_j f(\sigma_j)]$$

for all  $k = 0, \dots, d+1$ .

**Definition 2.10.** Let  $\delta$  be a metric on  $\mathbb{R}^d$ . The Vietoris-Rips complex at scale  $\epsilon \geq 0$  on  $X \subseteq \mathbb{R}^d$  is the abstract simplicial complex

$$\text{VR}_\epsilon(X) := \{[x_0, \dots, x_k] : \delta(x_i, x_j) \leq 2\epsilon, x_0, \dots, x_k \in X, k = 0, \dots, n\}.$$

If  $X$  is a dense enough sample from a manifold  $M \subseteq \mathbb{R}^d$ , then  $\text{VR}_\epsilon(X)$  can recover in some sense the true topology of  $M$ .

- At scale  $\epsilon = 0$ , then  $\text{VR}_0(X) = \{[x] : x \in X\}$ , that is  $\text{VR}_0$  overfits the data  $X$ .
- As  $\epsilon \rightarrow \infty$  all the points of  $X$  become vertices of a single  $|X|$ -dimensional simplex, giving a contractible topological space.

A persistence barcode is an interval  $[\epsilon, \epsilon']$  where at the left end-point a new feature appears (is born) and at the right end-point the feature disappears (or dies). The length of the interval,  $\epsilon' - \epsilon$ , is the persistence of that feature. Computing persistence barcodes for homology groups is known as persistent homology. In the following consider the finite set of scales  $\epsilon_0 < \dots < \epsilon_m$  and the simplicial complexes  $K_j = \text{VR}(X, \epsilon_j)$ .

**Definition 2.11.** *The filtration of simplicial complexes is the chain of nested simplicial complexes*

$$K_0 \subseteq \dots \subseteq K_m,$$

with the inclusion maps  $f_j : K_j \hookrightarrow K_{j+1}$  for  $j = 0, \dots, m-1$ .

**Remark 2.12.** *The index  $j$  is referred to as time. If a homology class is in  $H_k(K_{j+1})$  is not in the image  $H_k(K_j)$ , then the class is born at time  $j+1$ . If simplices of different homology classes in  $H_k(K_i)$  are mapped to the same homology class of  $H_k(K_j)$  for  $i < j$  then one class is said to have died, and the other is said to have persisted from time  $i$  to  $j$ .*

To identify classes that persist from time  $j$  to  $j+p$  it is sufficient to consider the  $p$ -persistent  $k^{\text{th}}$  homology group

$$H_k^{j,p} = Z_k^j / \left( B_k^{j+p} \cap Z_k^j \right).$$

### 3 Binary Classification

For binary classification, the aim is to classify two different probability distributions supported on disjoint manifolds  $M_a, M_b \subseteq \mathbb{R}^d$ . Suppose there exists a classifier with zero prediction error. Sample a large finite set of points  $T \subseteq M_a \cup M_b$  uniformly and densely and deduce the Betti numbers of  $M_a$  and  $M_b$ . Consider the feed-forward neural network  $v : \mathbb{R}^d \rightarrow [0, 1]$  given by

$$v = s \circ f_l \circ \dots \circ f_1,$$

where  $f_j : \mathbb{R}^{n_j} \rightarrow \mathbb{R}^{n_{j+1}}$  is a network layer involving an affine map  $\rho_j : \mathbb{R}^{n_j} \rightarrow \mathbb{R}^{n_{j+1}}, x \mapsto A_j x + b_j$  composed with an activation function  $\sigma : \mathbb{R}^{n_{j+1}} \rightarrow \mathbb{R}^{n_{j+1}}$ . The function  $s : \mathbb{R}^{n_{l+1}} \rightarrow [0, 1]$  is the score function. Let

$$v_j = f_j \circ \dots \circ f_1 \text{ and } v = s \circ v_l.$$

The output of the network is interpreted to be the probability that the input is on  $M_a$ . A well-trained network correctly classifies all  $x \in T$ . In reality, the topologies  $M_a$  and  $M_b$  may be intertwined in a complicated manner. Using the theory built up in the previous section it can be shown experimentally that the neural network unravels the decision boundary to opposite ends of  $[0, 1]$  for effective classification.

### 4 Architectural Implications on Data Topology

Every network trained with differing activation functions shows a decrease in  $\beta_0$  across the layers of the network. With  $\tanh$  activation the reduction is less effective, whereas ReLU demonstrated the largest decay due to the fact that ReLU is a non-homeomorphic activation function.

A Bottleneck layer forces large topological changes, and a narrow network changes topology faster than a wider one. Reducing the depth of a constant-width network makes it increasingly difficult to train the network to high accuracy. As the burden of changing topology does not spread evenly across the layers, but instead is concentrated at the final layers. Which reaches the saturation limits of the layers.

## References

- [1] Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. “Topology of deep neural networks”. In: *CoRR* (2020).