On Characterizing the Capacity of Neural Networks Using Algebraic Topology

William H. Guss Ruslan Salakhutdinov

Machine Learning Department, Carnegie Mellon University

Abstract

The learnability of different neural architectures can be characterized directly by computable measures of data complexity. In this paper, we reframe the problem of architecture selection as understanding how data determines the most expressive and generalizable architectures suited to that data, beyond inductive bias. After suggesting algebraic topology as a measure for data complexity, we show that the power of a network to express the topological complexity of a dataset in its decision boundary is a strictly limiting factor in its ability to generalize. We then provide the first empirical characterization of the topological capacity of neural networks. Phenomena therein allows us to connect existing theory to empirically driven conjectures on the choice of architectures for a single hidden layer neural networks.

The Problem of Architecture Selection

Deep learning has been extremely successful, in part due to the elimination of *feature engineering*.

Research focus has shifted:
best features → best *architectures*.

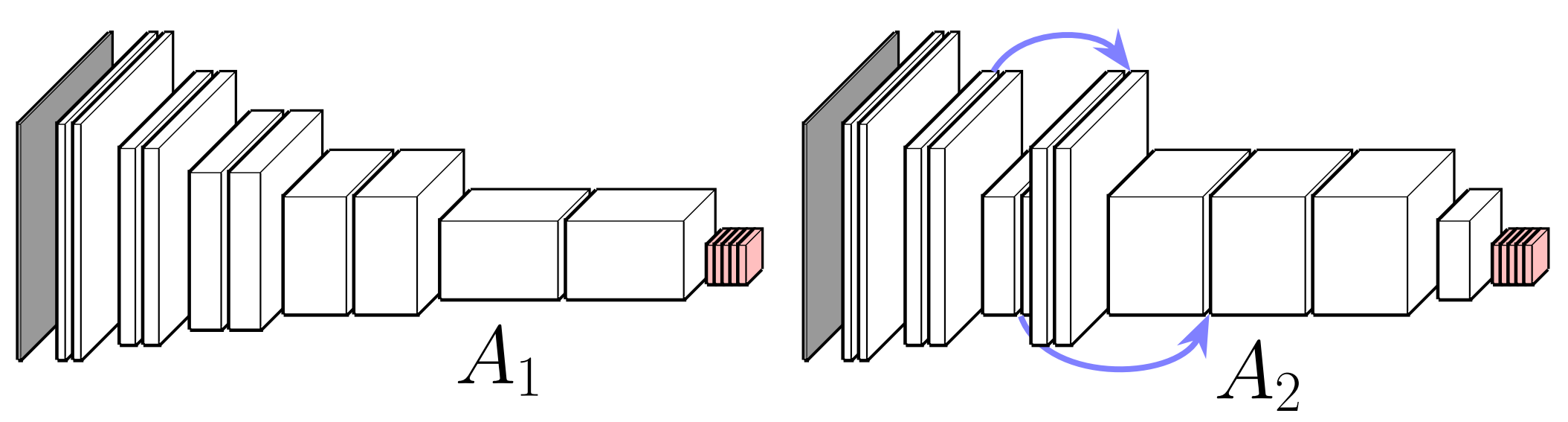

Figure 1: Examples of two competing convolutional architecture A_1 and A_2.

In computer vision, for example, a large body of work ([SZ14, SLJ+14, HZRS15], etc.) focuses on improving the initial architectural choices of [KSH12].

Despite the success of this approach, **there are still not general principles for choosing architectures in arbitrary settings.**

- **Neural architecture search** yields expressive and powerful architectures at the cost of interpratibility. [SV16, FBB+17, ZL17])
- **Expressivity theory** can only be used to determine an architecture in practice if it is understood how expressive a model need be in order to solve a problem. [RPK+16, DFS16, Gus16]
- **Data-first architecture selection:** develop some objective measure of data complexity, and then characterize neural architectures by their ability to learn subject to that complexity.

[B+14] Monica Bianchini et al.
On the complexity of shallow and deep neural network classifiers.
In *ESANN*, 2014.

[DFS16] Amit Daniely, Roy Frostig, and Yoram Singer.
Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity.
In *Advances In Neural Information Processing Systems*, pages 2253–2261, 2016.

[FBB+17] Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A. Rusu, Alexander Pritzel, and Daan Wierstra.
Pathnet: Evolution channels gradient descent in super neural networks.
CoRR, abs/1701.08734, 2017.

[Gus16] William H Guss.
Deep function machines: Generalized neural networks for topological layer expression.
arXiv preprint arXiv:1612.04799, 2016.

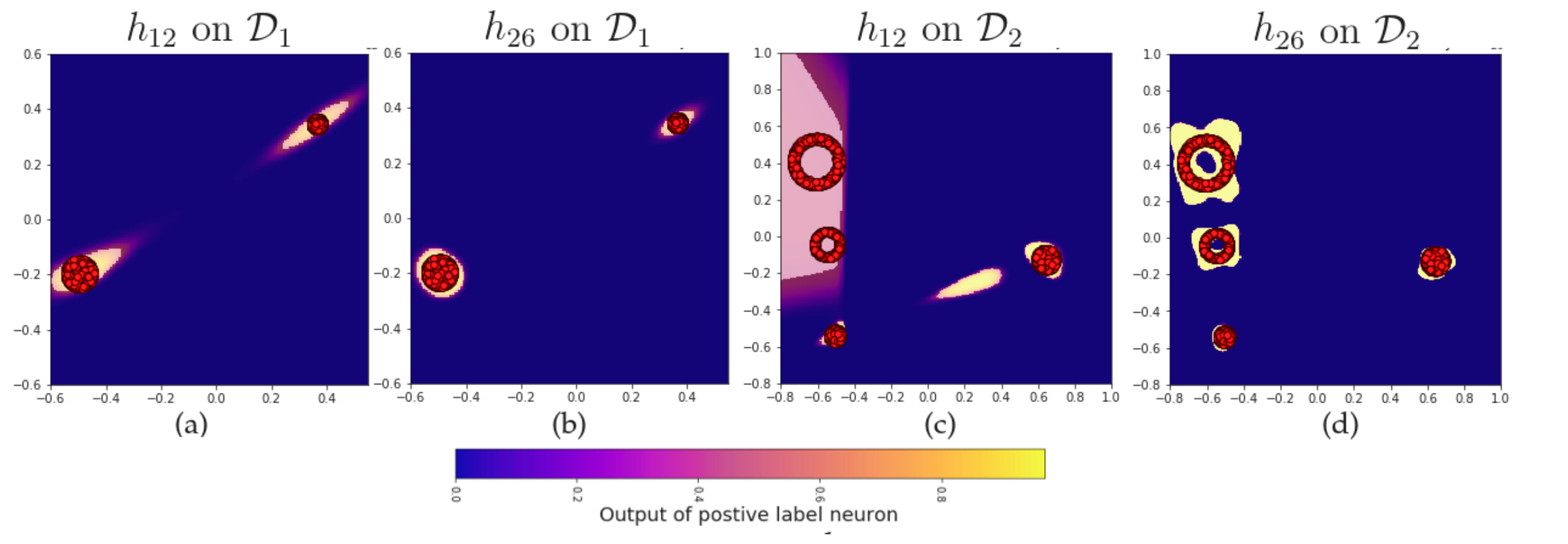

Figure 2: The positive label outputs of single hidden layer neural networks, h_{12} and h_{26}, of 2 inputs with 12 and 26 hidden units respectively after training on datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ with positive examples in red. Highlighted regions of the output constitute the positive decision region.

Background

Topology characterizes shapes and sets by their *connectivity*.

Def (Homeomorphism). Informally, we say that two topological spaces A and B are *equivalent* ($A \cong B$) if there is a continuous function $f : A \to B$ that has an inverse f^{-1} that is also continuous.

Topology differentiates sets in a meaningful way, discarding irrelevant properties like rotation, translation, curvature, etc.

Figure 3: A continuous deformation of a coffee cup into a donut, showing that both are topologically equivalent ([KYD14]).

Analysis of Figure 2.

$\mathcal{D}_1 \not\cong \mathcal{D}_2$, and h_{12} cannot express decision boundaries with the topology of $\mathcal{D}_2$.

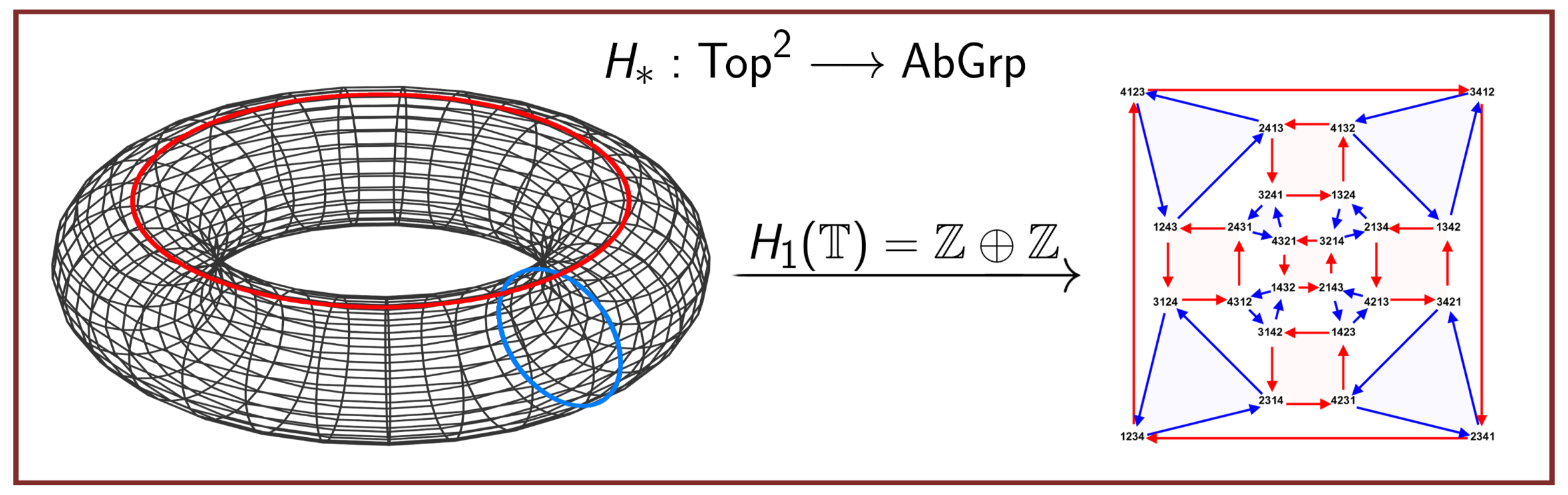
$H_* : \mathrm{Top}^2 \longrightarrow \mathrm{AbGrp}$
$H_1(\mathbb{T}) = \mathbb{Z} \oplus \mathbb{Z}$

Figure 4: An illustration of the core philosophy of algebraic topology: 'functoraly' reduce hard problems in topology to easy ones in group theory.

Def (Homology). If X is a topological space, then $H_n(X) = \mathbb{Z}^{\beta_n}$ is called the nth *homology group* of X if the power β_n is the number of 'holes' of dimension n in X. We call $\beta_n(X)$ the nth Betti number of X. Finally, the homology of X is defined as $H(X) = \{H_n(X)\}_{n=0}^{\infty}$.

Analysis of Figure 2.

$$H_0(\mathcal{D}_1) = \mathbb{Z}^2 \quad H_1(\mathcal{D}_1) = \{0\} \quad H_2(\mathcal{D}_1) = \{0\}$$

$$H_0(\mathcal{D}_2) = \mathbb{Z}^4 \quad H_1(\mathcal{D}_2) = \mathbb{Z}^2 \quad H_2(\mathcal{D}_2) = \{0\}$$

$H(\mathcal{D}_1) \leq H(\mathcal{D}_2)$ and $\mathcal{D}_2$ requires more complex architectures

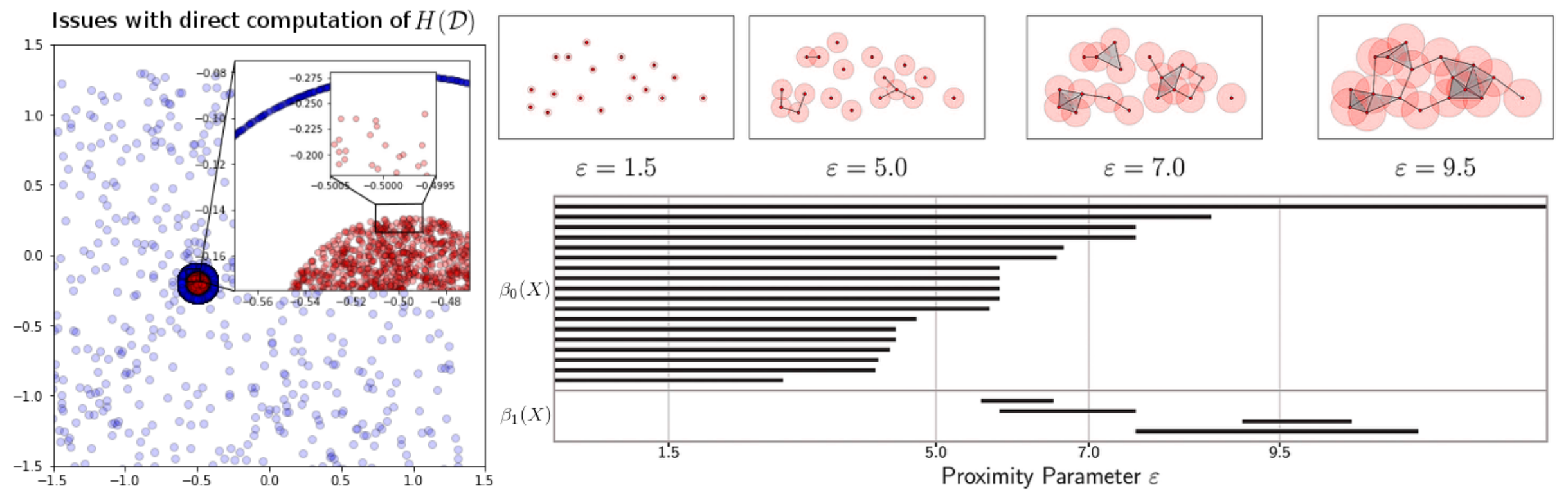

Figure 5: Left: The local disconnectedness of datasets prevents direct computation of their homology. Right: An illustration of computing persistent homology on a collection of points ([TZH15])

Basis for homological characterization

Homology is a stringent measure for characterizing architectures:

Let $\mathcal{D}$ be a class of data drawn from a topological manifold $\mathcal{M} \subset X$. Let $H_S(f)$ be the homology of the support of f, i.e. $H_S(f) = H(\{x : f(x) > 0\})$.

Theorem (The Homological Principle of Generalization). If for all $f \in \mathcal{F}$, $H_S(f) \neq H(\mathcal{M})$, then there exists $A \subset X$ such that f missclassifies A.

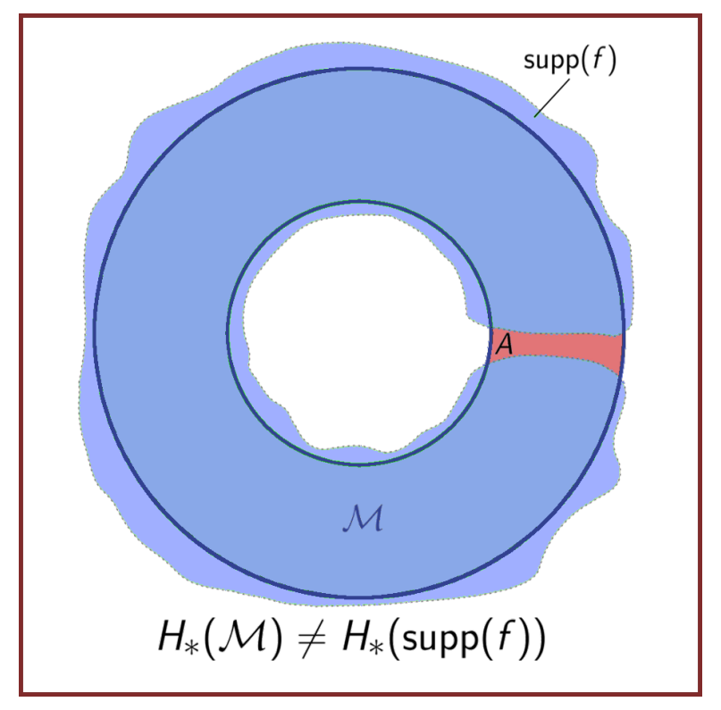

Figure 6: When homology cannot be expressed, there exists a misclassified set.

Given a dataset $\mathcal{M}$, for which architectures A does there exist a neural network $f \in \mathcal{F}_A$ such that $H_S(f) = H(\mathcal{M})$?

Empirical expressivity of neural networks

We characterize low dimensional neural networks by generating random synthetic datasets of a given homological complexity and measuring the homology of super-level sets over the course of training.

Topological Phase Transitions

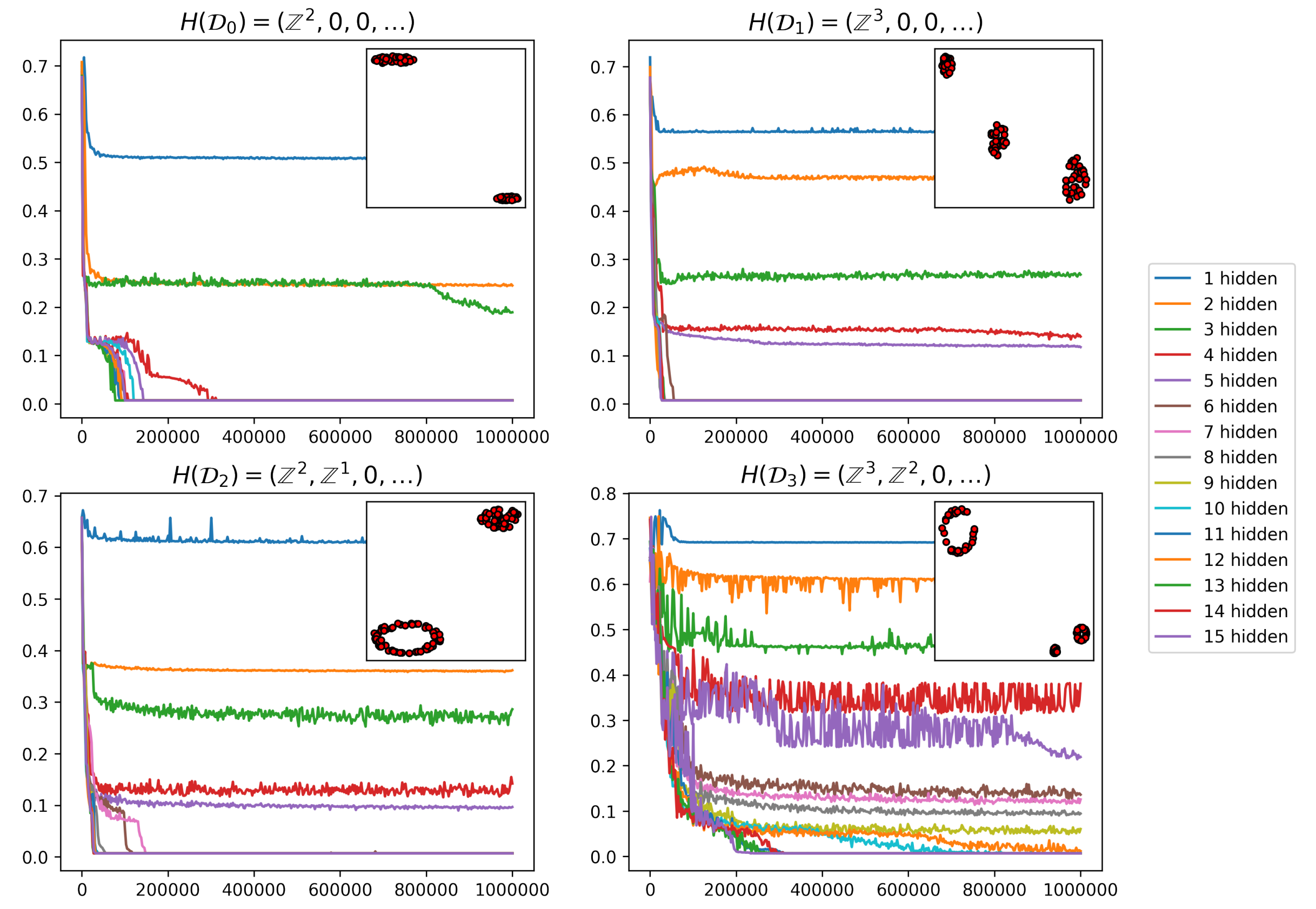

Figure 7: Topological phase transitions in low dimensional neural networks as the homological complexity of the data increases. The upper right corner of each plot is a dataset on which the neural networks of increasing hidden dimension are trained.

Topological Stratification

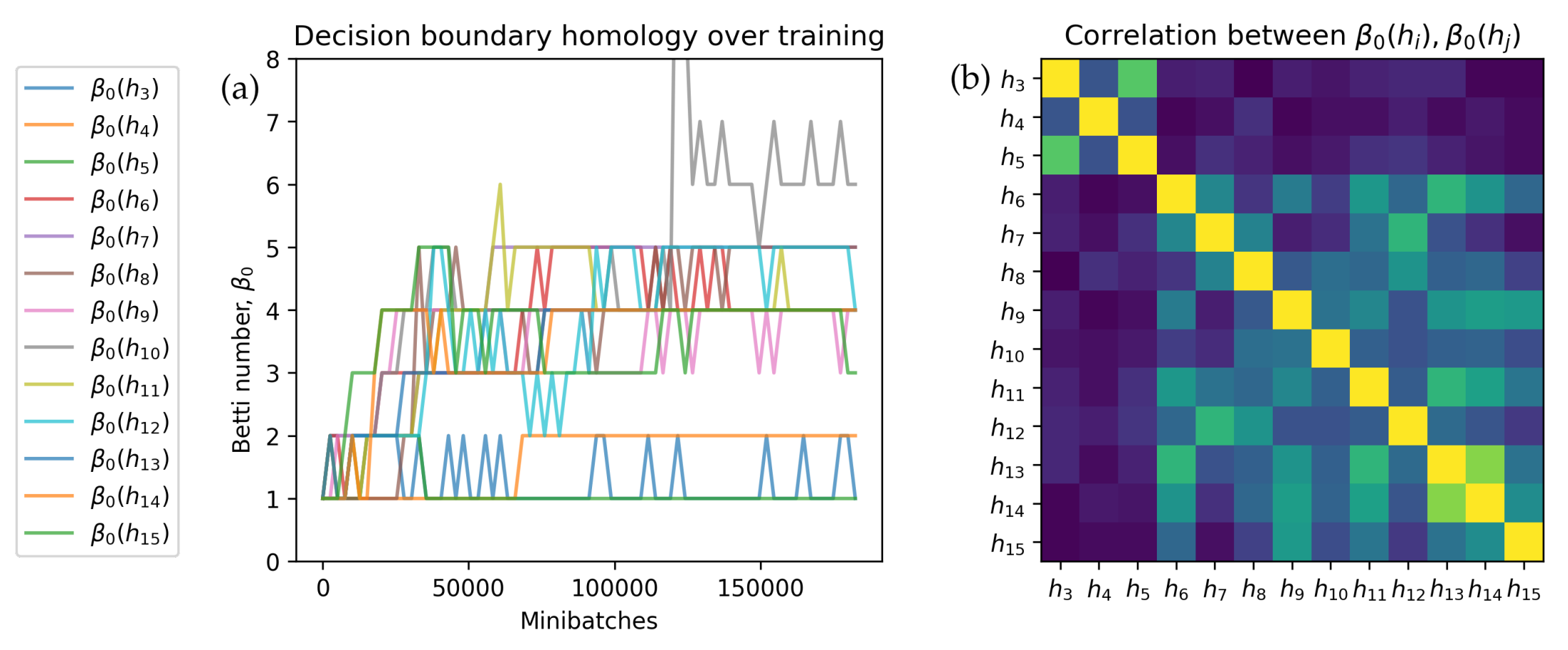

Figure 8: An example of topological stratification for single hidden layer networks. (a) The number of connected components in the decision regions during training. (b) Correlation of Betti numbers.

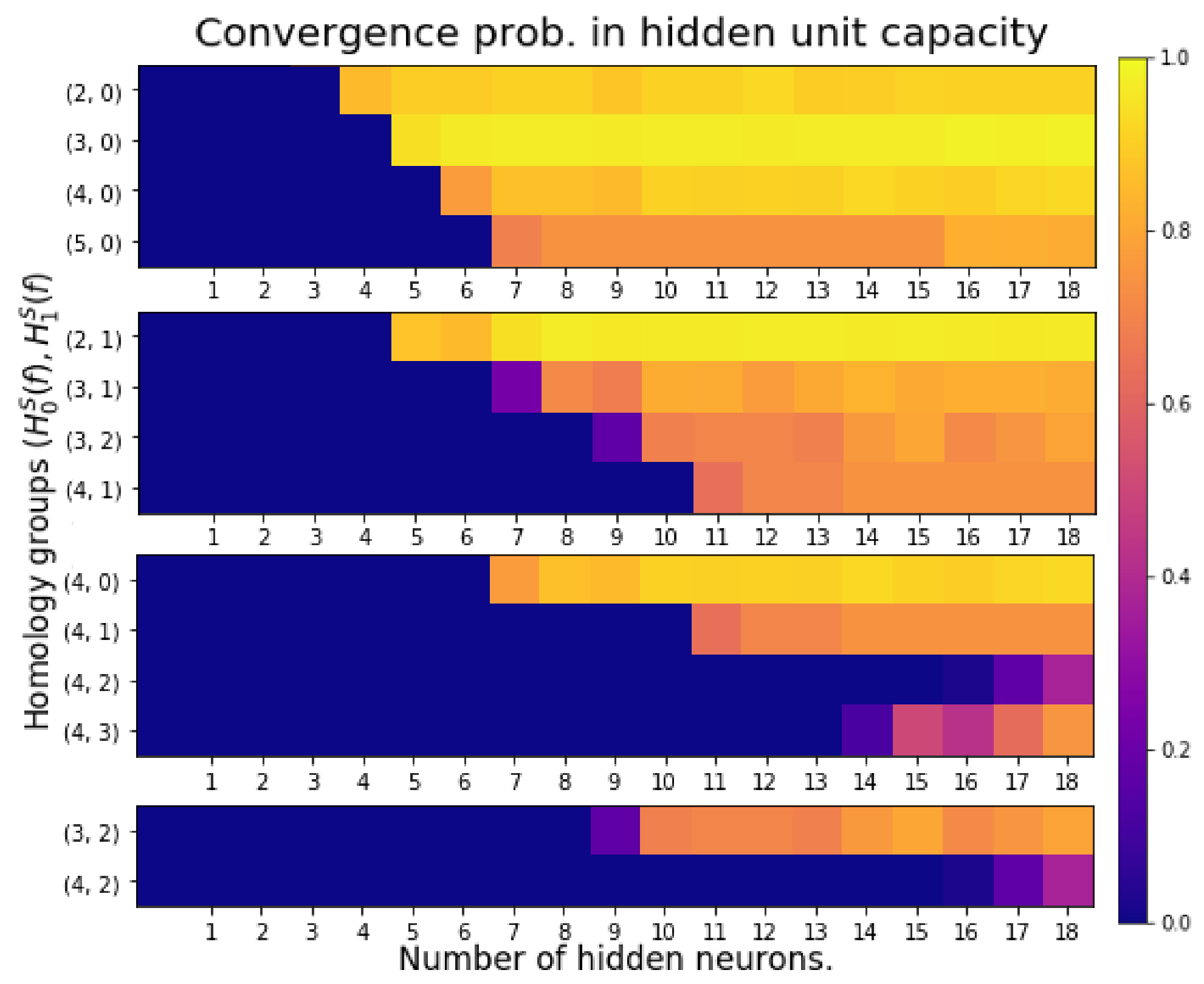

Figure 9: Several different views of the probability of converging to zero-error for single hidden layer neural networks on datasets with different homological complexities.

Topology of Real Data

Real data is homologically rich:

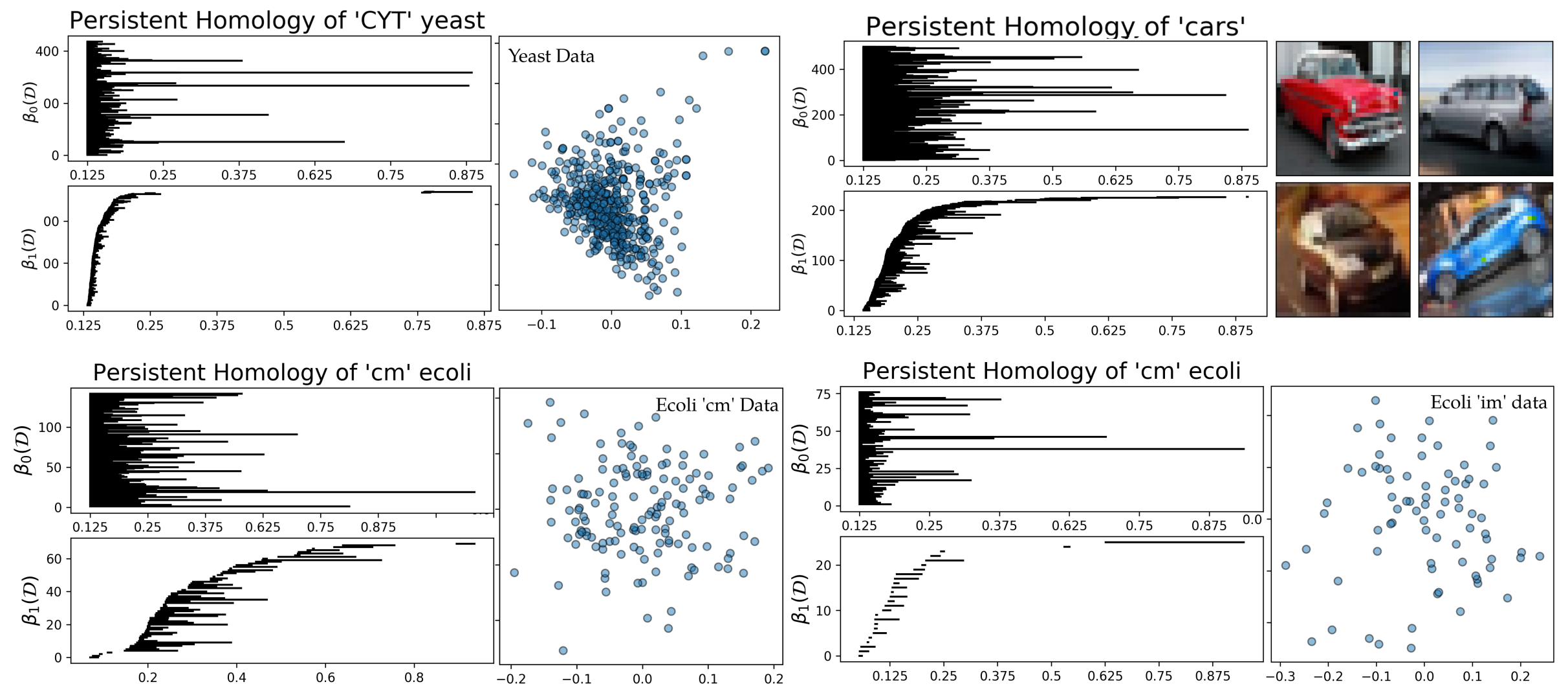

Figure 10: The persistent homology barcodes of classes in the CIFAR-10 and UCI Protein Localization Datasets. Left: The bardcode for dimension 0 and 1 of the 'CYT' class along side its local linear embedding into $\mathbb{R}^2$. Right: The barcode for the dimensions 0 and 1 for the 'cars' class along side different samples thereof in CIFAR-10. Note how different orientations are shown.

Towards a complete neural homology theory

The initial work of [B+14] gives bounds on $\sum \beta_n$, but exact characterization is crucial to architecture search.

We've progress towards exact characterization using the Mayer-Viatoris sequence:

$$\ldots \xrightarrow{\partial_*} H_p(A \cap B) \xrightarrow{i_*^A \oplus i_*^B} H_p(A) \oplus H_p(B) \xrightarrow{j_*^B - j_*^B} H_p(A \cup B) \xrightarrow{\partial_*} H_{p-1}(A \cap B) \longrightarrow \ldots$$

Theorem(s). Let h_k be a SHLN of $n = 2$ inputs and k hidden units with monotonic activation. Denote w_i as the ith weight vector.

$$H_0(h_1) = \chi_{C_2^1(w_1,\beta)}\mathbb{Z} \trianglelefteq \mathbb{Z}^1$$

$$H_0(h_2) = \chi_{C_2^2(w_1,w_2,\beta)}\mathbb{Z} \oplus \mathbb{Z} \trianglelefteq \mathbb{Z}^2$$

$$H_0(h_3) = \frac{\chi_{C_2^2(w_1,w_2,\beta)}\mathbb{Z} \oplus \mathbb{Z} \oplus \mathbb{Z}}{\dfrac{\left[\chi_{C_2^2(w_1,w_2,\beta)}\mathbb{Z}\right] \oplus \left[\chi_{C_2^2(w_2,w_3,\beta)}\mathbb{Z}\right]}{\chi_{\neg \bigvee_{i \neq j} C_2^2(w_i,w_j) \vee C_3^\Upsilon(w_1,w_2,w_3)}\mathbb{Z}}} \trianglelefteq \mathbb{Z}^3$$

$$H_1(h_3) = \frac{\dfrac{\left[\chi_{C_2^2(w_1,w_2,\beta)}\mathbb{Z}\right] \oplus \left[\chi_{C_2^2(w_2,w_3,\beta)}\mathbb{Z}\right]}{\chi_{\neg \bigvee_{i \neq j} C_2^2(w_i,w_j) \vee C_3^\Upsilon(w_1,w_2,w_3)}\mathbb{Z}}}{ker(i_*^{h_2} \oplus i_*^{h_1})} \trianglelefteq \mathbb{Z}^1$$