Persistent Homology Dimension Does Not Measure Generalisation

Charlie Tan

Abstract

Neural network optimisation trajectories have been shown to possess fractal structure. Recent work connects the generalisation performance of a neural network to the fractal dimension of its trajectory. Persistent homology dimension is an approximation of fractal dimension employing topological data analysis, and has been applied as a measure of generalisation. In this work, persistent homology dimension is demonstrated to fail to measure generalisation at high learning rates, or when using adversarial initialisation. Despite this, persistent homology dimension is observed to correlate with test accuracy for many experimental configurations; remarkably succeeding to measure model-wise double descent.

1 Introduction

The widespread empirical success of deep learning enjoys limited theoretical support. Of particular interest is the implicit bias of neural networks towards generalisation despite over-parameterisation Zhang et al. (2017). Measures from statistical learning theory, such as Rademacher complexity Bartlett and Mendelson (2003) and VC-Dimension Hastie et al. (2009), indicate that without explicit regularisation over-parameterised models will generalise poorly. In contrast, neural networks are able to generalise strongly despite having sufficient capacity to simply memorise their training data Liu et al. (2021). Remarkably, neural networks often exhibit *improved* generalisation for increases in capacity Nakkiran et al. (2019); novel learning theory is thus required.

Generalisation is typically attributed to the implicit bias of gradient-based optimisation. A number of works have considered the geometry of generalising solutions within parameter space, and the bias of optimisation methods towards such solutions Garipov et al. (2018); He et al. (2019), Further works analyse optimisation trajectories as discretisations of stochastic differential equations, with Smith et al. (2021) defining a modified loss function on which finite step-size stochastic gradient descent (SGD) follows gradient flow. Ultimately, a goal of deep learning theory research is to define generalisation bounds for given experimental configurations Valle-Pérez and Louis (2020).

Fractals are shapes typified by self-similarity across scales, where 'zooming' into the shape results in similar structures reappearing at increasingly small scales. Fractal dimension describes the rate at which a fractal scales with respect to its basis; the self-similarity being described by a non-integer rate of scaling. Recently, Simsekli et al. (2020) demonstrated neural network optimisation trajectories to possess random-fractal structure, and proposed a generalisation bound based on fractal dimension. However, this work assumed both topological and statistical conditions on the optimisation trajectory. The work of Birdal et al. (2021) employs persistent homology dimension ($\dim_{\mathrm{PH}}$) Adams et al. (2020), a measure of fractal dimension from topological data analysis, to relax these assumptions. They propose an efficient procedure for estimating persistent homology dimension, and apply this both as a measure of generalisation and a scheme for explicit regularisation. We note another proposed topological generalisation measure, named neural persistence, that considers only the final iterate (not a trajectory) and is limited to fully-connected networks Rieck et al. (2019).

This project is a primarily empirical investigation of persistent homology dimension as a generalisation measure, in which robustness and failure modes are explored in a wider range of experiments to those considered by Birdal et al. (2021). A central result of this project is evidence of persistent homology dimension failing to correlate with generalisation, induced by both large learning rates or adversarial initialisation Zhang et al. (2017). Whilst multiple failure modes are identified, persistent homology dimension is found to correlate remarkably well with test accuracy throughout several experimental configurations, inviting further analysis of this relationship.

The full contributions of this work are:

1. reproduction of results, confirming the correlation between $\dim_{\mathrm{PH}}$ and generalisation at the range of learning rates considered by Birdal et al. (2021).
2. extended learning rate experiment, in which $\dim_{\mathrm{PH}}$ is demonstrated to fail to correlate with generalisation at large learning rates.
3. novel experiment in which $\dim_{\mathrm{PH}}$ is demonstrated unable to identify a poorly generalising model, trained from an adversarial initialisation Zhang et al. (2017).
4. novel experiment in which $\dim_{\mathrm{PH}}$, along with test accuracy, are demonstrated to undergo model-wise double descent Nakkiran et al. (2019).

2 Preliminaries and Technical Background

2.1 Homology Groups and Betti Numbers

$$\begin{aligned} H_0(S^1) &= \mathbb{Z} \\ H_1(S^1) &= \mathbb{Z} \end{aligned} \quad \longrightarrow \quad \begin{aligned} \beta_0(S^1) &= 1 \\ \beta_1(S^1) &= 1 \end{aligned}$$

$$\begin{aligned} H_0(B^2) &= \mathbb{Z} \\ H_1(B^2) &= \{0\} \end{aligned} \quad \longrightarrow \quad \begin{aligned} \beta_0(B^2) &= 1 \\ \beta_1(B^2) &= 0 \end{aligned}$$

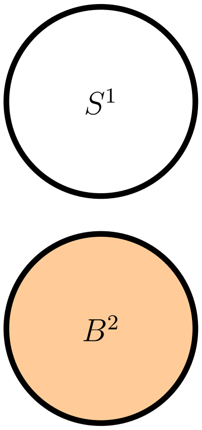

Figure 1: The circle S^1 and solid disk B^2 differ in first homology group. The kth Betti number β_k is the rank of the kth homology group H_k, and corresponds to the number of k-dimensional holes present. Both the circle and disk have a single connected component (a zero-dimensional hole by convention) and thus $\beta_0(S^1) = \beta_0(B^2) = 1$. However, the circle has a one-dimensional hole whereas the disk does not; accordingly $\beta_1(S^1) = 1 \neq \beta_1(B^2) = 0$. Here $\{0\}$ denotes the trivial group.

Algebraic topology is the study of topological spaces using abstract algebra. Topological spaces can be classified using topological invariants, properties invariant under homeomorphisms. Holes are one such invariant; a disk can be shrunk into a point whereas a circle can not due to the (one-dimensional) hole present. The disk is therefore homeomorphic to a point, whilst the circle is not, and these spaces are thus topologically distinct. Homology formalises the notion of holes, as displayed in Figure 1, a brief introduction follows Birdal et al. (2021); Wikipedia (2023).

Let $\mathcal{W}$ be a topological space. A chain complex $C(\mathcal{W})$ on $\mathcal{W}$ is a sequence of (abelian) chain groups C_k connected by boundary operators $\delta_k : C_k \mapsto C_{k-1}$. Here $k \in \mathbb{N}$ corresponds to dimension.

$$C(\mathcal{W}) = \cdots\ C_{k+1} \xrightarrow{\delta_{k+1}} C_k \xrightarrow{\delta_k} C_{k-1} \cdots$$

The composition of consecutive boundary operators is necessarily the trivial group $\delta_{k+1} \circ \delta_k \equiv \{0\}$. The elements of $\mathrm{im}(\delta_{k+1})$ are referred to as boundaries, and those of $\ker(\delta_k)$ as cycles. The kth homology group of $\mathcal{W}$ is defined as the quotient group of cycles modulo boundaries.

$$H_k(\mathcal{W}) := \ker \delta_k / \operatorname{im} \delta_{k+1}$$

Elements of homology groups are equivalence classes of cycles that differ by boundary. The kth **Betti number** β_k is the rank of the kth homology group H_k and defines the number of k-dimensional holes in $\mathcal{W}$, $\beta_k(\mathcal{W}) = \mathrm{rank}\,(H_k(\mathcal{W}))$. We note that by convention, zero-dimensional holes refer to connected components. Betti numbers are a topological invariant and can be used to classify spaces.

2.2 Topological Data Analysis and Persistent Homology

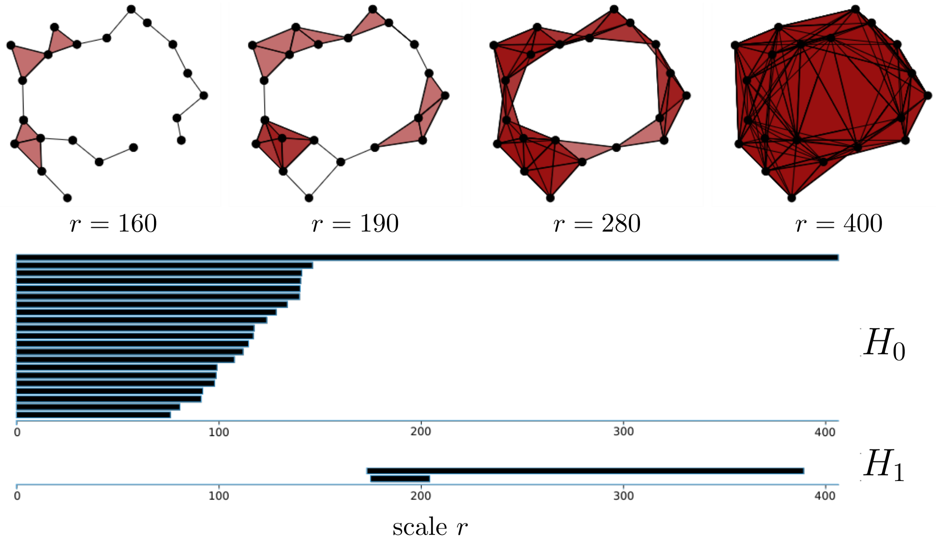

Figure 2: Adapted from Adams (2021). **A filtration is a sequence of Vietoris-Rips complexes at increasing scale**. **Above:** Vietoris-Rips complexes at four scales. **Below:** persistence barcode resulting from filtration over $r \in [0, 400]$. The persistence barcode uniquely tracks each k-dimensional hole γ from birth to death. When $r = 0$ every point is an isolated connected component (a zero-dimensional hole), as seen by the large number of elements in H_0 at this scale. As r increases these components are connected by lines, and by $r = 160$ there is only a single connected component remaining. At $r = 185$ two one-dimensional holes are born, ones of these is 'noisy' and dies soon after $r = 200$, however the other is more representative of the data topology and 'persists' much longer until $r = 390$. Only H_0 and H_1 displayed.

Topological data analysis seeks insight from datasets using techniques from topology. In this framework, data are considered as samples from a topological space; the objective is to understand the topology of this space from finite samples Chazal and Michel (2021). A simplicial complex (a high dimensional generalisation of a graph) is constructed from the data, and the simplicial homology analysed in order to understand the topology of the dataset. Constructing a simplicial complex requires defining a scale, this determines how 'close' data must be to be considered connected. **Persistent homology** is a fundamental technique in topological data analysis that summarises homology over a range of scales Edelsbrunner et al. (2000); Zomorodian and Carlsson (2004). Intuitively 'important' topological, features will persist over many scales, whilst 'noisy' ones will not.

While there are multiple methods for constructing a simplicial complex from a point cloud, persistent homology dimension employs the **Vietoris-Rips complex**. The notion of 'a range of scales' is encoded by a **filtration**.

Definition 1 Vietoris-Rips Complex: *For W a point cloud in a metric space, a Vietoris-Rips complex $\mathrm{VR}_r(W)$ at scale r is constructed using the distances between points. Intuitively, any set of k points that can be contained within a ball of radius r are considered connected by a $k-1$ dimensional simplex. For example, $k = 2$ points within an r-ball are connected by a 1-dimensional simplex (line), and $k = 3$ points by a 2-dimensional simplex (triangle).*

Definition 2 Filtration: *For a given point cloud W in a metric space, different Vietoris-Rips complexes will result at different scales r. A filtration $\mathrm{VR}(W)$ is a sequence of Vietoris-Rips complexes at increasing scale.*

Across a filtration, k-dimensional holes are created (born) or filled (died); the homology varies with scale. The birth and death of each k-dimensional hole γ is tracked and represented using a **persistence barcode**. We define birth (γ) and death (γ) as the scales at which the hole appears and disappears, and the lifetime length $|I(\gamma)|$ as death (γ) − birth (γ) Simsekli et al. (2020). At any given scale r the number of k-dimensional holes present is precisely the Betti number of the Vietoris-Rips complex at this scale. We note that computing persistent homology dimension only requires the zeroth dimension homology, corresponds to the connected components. Samples from an example Vietoris-Rips filtration and the corresponding persistence barcode are provided in Figure 2.

We may now define the **persistent homology dimension** $\dim_{\mathrm{PH}}$, an estimator for fractal dimension using topological data analysis Adams et al. (2020). We provide the descriptive definition of $\dim_{\mathrm{PH}}$ from Birdal et al. (2021).

Definition 3 Persistent Homology Dimension: *Birdal et al. (2021): For a finite set W regarded as samples from a topological space $\mathcal{W}$, the persistent homology dimension $\dim_{\mathrm{PH}}$ is the smallest α for which the α-weighted lifetime sum E^i_α is uniformly bounded for all finite subsets of $\mathcal{W}$:*

$$E^i_\alpha(W) = \sum_{\gamma \in \mathrm{PH}_i(VR(W))} |I(\gamma)|^\alpha$$

2.3 Persistent Homology Dimension as a Measure of Generalisation

The result that non-convex stochastic optimisation trajectories exhibit fractal-like properties was established by Simsekli et al. (2020). However, in order to invoke the Hausdorff dimension, both topological and statistical conditions were assumed on the optimisation trajectory. The work of Birdal et al. (2021) relaxed these assumptions using techniques from topological data analysis.

We follow Birdal et al. (2021) in defining a standard supervised learning setting; given a data space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ generated by unknown distribution $\mathcal{D}$ we seek to learn a mapping from features $x \in \mathcal{X}$ to labels $y \in \mathcal{Y}$. With $\mathcal{D}$ unknown, we are required to achieve this using Q training samples $S = \{z_i\}_{i=1}^Q$ drawn independently and identically from $\mathcal{D}$. We define $w \in \mathbb{R}^d$ to be the weight vector of a neural network, and $\mathcal{W}$ to be the set containing the optimisation trajectory this weight vector undergoes during T steps of training $\mathcal{W} = \{w_i\}_{i=1}^T$. To evaluate the quality of a weight vector $w \in \mathcal{W}$ for a given

Algorithm 1 Birdal et al. (2021): Computation of $\dim_{\mathrm{PH}}$.

```
input: The set of iterates $W = \{w_i\}_{i=1}^{K}$, smallest sample size $n_{\min}$, skip step $\Delta$, and $\alpha$
output: $\dim_{\mathrm{PH}} W$
$n \leftarrow n_{\min}$, $E \leftarrow []$
while $n \leq K$ do
    $W_n \leftarrow \mathrm{sample}\,(W, n)$                                   ▷ randomly sample $n$ points from $W$
    $\mathcal{W}_n \leftarrow \mathrm{VR}(W_n)$                                ▷ Vietoris-Rips filtration on samples
    $E[i] \leftarrow E_\alpha(\mathcal{W}_n) \triangleq \sum_{\gamma \in \mathrm{PH}_0(\mathcal{W}_n)} |I(\gamma)|^\alpha$   ▷ compute lifetime sums
    $n \leftarrow n + \Delta$                                                  ▷ step $n$
end while
$m, b \leftarrow \mathrm{fitline}\,(\log\,(n_{\min} : \Delta : K), \log\,(E))$   ▷ fit line, get gradient $m$ and bias $b$
$\dim_{\mathrm{PH}} W \leftarrow \frac{\alpha}{1-m}$
```

sample $z \in \mathcal{Z}$ we employ a loss function $\ell : \mathbb{R}^d \times \mathcal{Z} \mapsto \mathbb{R}_+$. Thus we define the population risk as $\mathcal{R}(w) \coloneqq \mathbb{E}_{\mathcal{Z}}[\ell(w, z)]$ and training risk as $\hat{\mathcal{R}}(w, S) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i)$, upon which generalisation error can be defined as $|\hat{\mathcal{R}}(w, s) - \mathcal{R}(w)|$.

With the problem setting defined, Theorem 1 is a central result of Birdal et al. (2021). This theorem closely resembles Theorem 1 of Simsekli et al. (2020), with a key difference being that persistent homology dimension $\dim_{\mathrm{PH}}$ replaces Hausdorff dimension $\dim_{\mathrm{H}}$. Defining M is beyond the scope of this report, but it is constant for a given network architecture and dataset. Theorem 1 implies that decreasing persistent homology dimension tightens the upper bound on generalisation error, suggesting its utility as a measure for generalisation.

Theorem 1 *Birdal et al. (2021): Let $\mathcal{W}$ be a bounded compact set. Assuming loss ℓ is bounded by B and L-Lipschitz continuous in w, for constant $M \geq 1$ and sufficiently large number of training samples Q.*

$$\sup_{w \in \mathcal{W}} |\hat{\mathcal{R}}(w, S) - \mathcal{R}(w)| \leq B\sqrt{\frac{[\dim_{\mathrm{PH}} \mathcal{W} + 1]\log^2(nL^2)}{Q} + \frac{\log\,(7M/\gamma)}{Q}} \tag{1}$$

Birdal et al. (2021) propose Algorithm 1 for the estimation of $\dim_{\mathrm{PH}} W$, where W is the final K weight iterations of an optimisation trajectory. The algorithm commences by taking random samples of size n from W, computing a Vietoris-Rips filtration on this sample, and then computing the zeroth dimension lifetimes sums from the filtration. Linear regression is then performed over the logarithms of n (the array of sample sizes used) and E (the array of resulting lifetime sums), providing a slope m, which is used to compute the estimate of $\dim_{\mathrm{PH}} W$. Two further inputs, $n_{\min}$ and Δ, define the minimum number of samples to be taken, and the step in sample size between filtrations; together with K these define the array of sample sizes n on which Vietoris-Rips filtrations are performed. Evidently, $\dim_{\mathrm{PH}} W$ is inversely proportional to the slope m. This slope represents the rate-of-change in (the logarithm of) lifetime sum with respect to (the logarithm of) the sample sizes. Taking the samples sizes defined by $n_{\min}$, Δ and K to be constant, we are therefore interested in the variation of lifetime sums over these constant values. We emphasise the random sampling taken; the filtration at each step n represents the persistent homology of a random subset of n elements from W.

3 Experimental Configuration

The following experiments explore persistent homology dimension in the following scenarios: (i) variation of learning rate (ii) variation of training length (iii) adversarial initialisation (iv) model-wise double descent.

All experiments are conducted with a 5-layer convolutional neural network, defined by Nakkiran et al. (2019) as 'standard CNN'. The model consists of 3×3 convolutional layers of widths $[c, 2c, 4c, 8c]$, where c is a width (channel) multiplier. In all experiments $c = 64$ unless otherwise stated. Each convolution is followed by an (optional) BatchNorm layer, a ReLU activation and a MaxPool operation with kernel = stride = $[1, 2, 2, 8]$. Two datasets are employed; CIFAR-10 and CIFAR-100, in no cases is data augmentation used. We train with cross-entropy loss using stochastic gradient descent (SGD) with no momentum nor weight decay, and a batch size of 128. No learning rate scheduling is employed unless explicitly stated.

We compute $\dim_{\mathrm{PH}}$ using Algorithm 1, implemented by the original authors Birdal et al. (2021). We use the default configuration of $K = 1000$ iterations prior to convergence / a predefined stopping epoch, with $n_{\min} = 200$, $\Delta = 50$. We follow Birdal et al. (2021) in defining convergence as the first epoch at which 100% training accuracy is achieved. Birdal et al. (2021) set $\alpha = 1$ for their experiments, commenting that it must be strictly lower than the intrinsic dimension of $\mathcal{W}$, which we additionally follow. Every experiment was repeated for three random seeds, each run is either individual presented, or as a mean and standard deviation.

4 Large Learning Rates

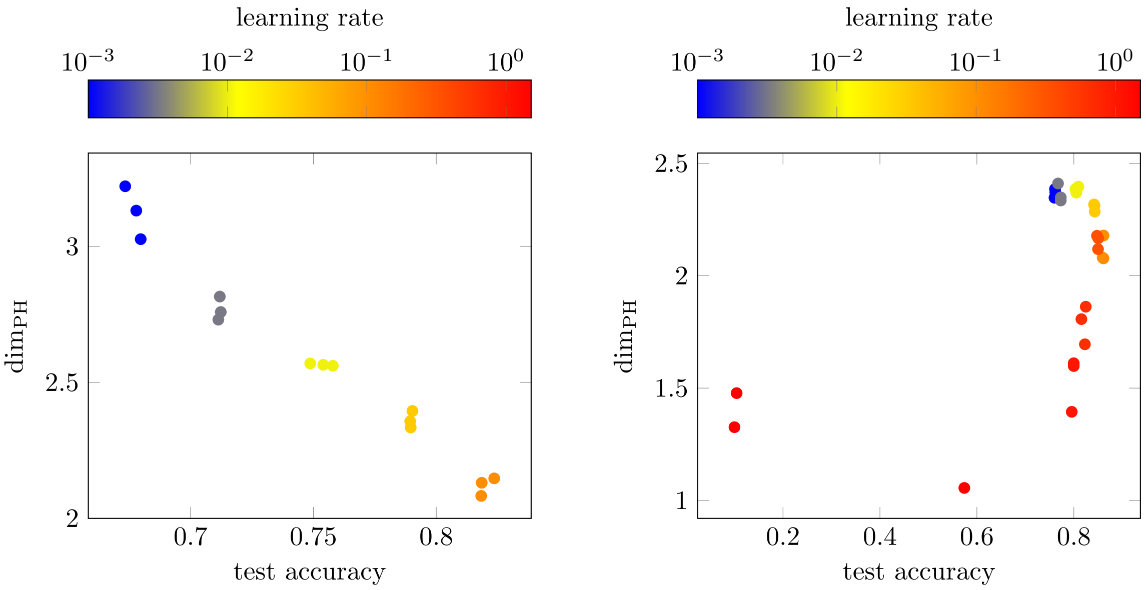

Figure 3: Persistent homology dimension $\dim_{\mathrm{PH}}$ correlates with test accuracy over a range of learning rates, but fails to do so at extremely large values. Left: without BatchNorm training is stable up to a learning rate of 10^{-1}, and $\dim_{\mathrm{PH}}$ correlates with test accuracy in this range. **Right: with BatchNorm** stable training is enabled above 10^{-1}, but at these higher values $\dim_{\mathrm{PH}}$ does not correlate with test accuracy. Note that since training is stopped once training accuracy = 1, generalisation error = 1 − test accuracy. Three seeds per learning rate.

Experiment details: We begin by reproducing the core experiments of Birdal et al. (2021), wherein a range of learning rates are employed to produce models of varying test accuracy. We repeat this experiment both with and without the use of BatchNorm. Following Birdal et al. (2021), we stop training once the training data has been perfectly fit (training accuracy reaches 100%). The results for these experiments are presented in Figure 3.

Discussion: whilst these results indicate that for ‘typical’ learning rates persistent homology dimension is indeed correlated with generalisation, they also illuminate the absence of this correlation at high learning rates. Birdal et al. (2021) only consider networks without BatchNorm, and the results of Figure 3 (left) are consistent with their hypothesis. However, using BatchNorm permits learning rates an order of magnitude higher whilst still converging to perfect training accuracy; it is at these high learning rates the correlation between persistent homology dimension and test accuracy fails. We note that for learning rates of up to 10^{-1}, the results using BatchNorm show the expected negative correlation, hence the presence of BatchNorm itself is not the failure mode. Importantly, we note a strong correlation between $\dim_{\mathrm{PH}}$ and learning rate, even at very high learning rates. It is therefore possible that a causal relationship exists between learning rate and $\dim_{\mathrm{PH}}$, and this is the source of the correlation with generalisation (since learning rates of up to 10^{-1} also correlate with generalisation).

5 Adversarial Pretraining

	Test Accuracy	$\dim_{\mathrm{PH}}$
Train on (Fixed) Random Labels	0.100 ±0.002	2.31 ±0.036
Train on (Fixed) Random Labels, then on True Labels	0.654 ±0.003	2.45 ±0.016
Train on True Labels Only	0.754 ±0.006	2.61 ±0.031

Table 1: Persistent homology dimension $\dim_{\mathrm{PH}}$ **incorrectly predicts higher test accuracy for a model trained from adversarial initialisation (middle row), than one trained from random initialisation (bottom row).** That test accuracy and $\dim_{\mathrm{PH}}$ are both lower when training an adversarially pretrained model (middle row) than training from random initialisation (bottom row) is in contrast to the negative correlation expected. Note that since training is stopped once training accuracy = 1, generalisation error = 1 − test accuracy. Mean of three seeds $\pm\sigma$.

Experiment details: we continue by employing adversarial pretraining as a method for generating poorly generalising models, known as ‘bad minima’ Zhang et al. (2017). With a possible spurious correlation between $\dim_{\mathrm{PH}}$ and learning rate, this experiment is defined to produce models of varying accuracy at the same learning rate. Following Zhang et al. (2017), we first pretrain using a randomised version of a given dataset, where random labels are generated prior to training and kept fixed throughout. This pretraining is executed until the model achieves train accuracy = 100%. In this case, since the data is unstructured the model is simply required to memorise the training samples. This pretrained model is then taken as the initialisation for another training phase, this time on the true dataset labels. The pretrained model can be thought of as an ‘adversarial initialisation’, from which SGD is highly unlikely to find a generalising solution. A measure of generalisation should be able to differentiate between a ‘bad minima’ and a ’good minima’ (trained only on true labels from random initialisation). CIFAR-10 was employed in these experiments with a learning rate of 10^{-2} for all training phases. No BatchNorm was used. The results for these experiments are presented in Table 1.

Discussion: these results demonstrate the failure of persistent homology dimension as a measure for generalisation at non-extreme learning rates. The configurations used in the middle and bottom rows of Table 1 differ only by the use of adversarial or random initialisation. The adversarially initialised model converges to a solution inferior to that of the random initialisation by 10% test accuracy; a successful measure of generalisation should be able to correctly predict such a large difference. However, the persistent homology dimension is also lower for the model that was first adversarially pretrained, which by the hypothesis of Birdal et al. (2021) is predictive of higher test accuracy. We further remark on the top row of Table 1, which are the results for the adversarial pretraining task itself. In this case we cannot directly relate $\dim_{\mathrm{PH}}$ to the low test accuracy, since the training dataset and test dataset are not drawn from the same distribution. However, we can note that $\dim_{\text{PH}}$ is *not* a measure of function simplicity, since the function required to memorise the randomised training samples will be highly complex. Furthermore, the low value of $\dim_{\text{PH}}$ is notable since this model will not generalise to *any* dataset; persistent homology dimension has failed to recognise memorisation.

6 Convergence Rate and Stability

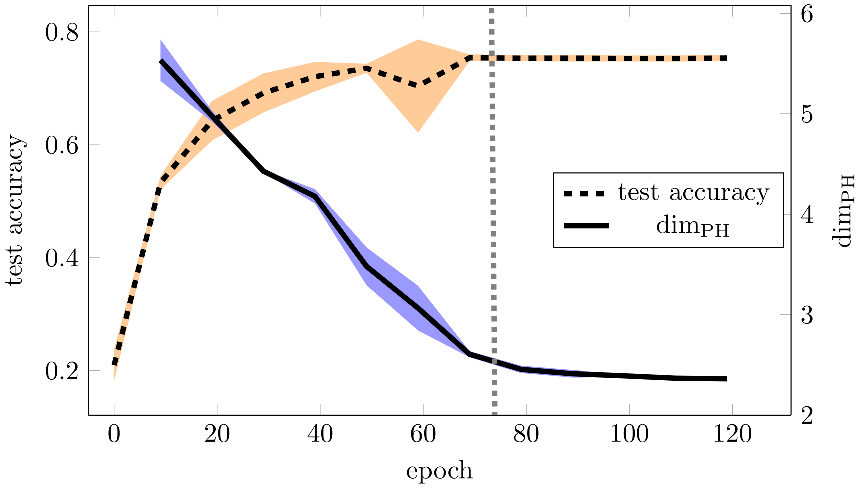

Figure 4: Persistent homology dimension $\dim_{\text{PH}}$ continues to converge after 100% training accuracy has been achieved. 100% training accuracy is achieved at epoch 75 (vertical line), at which point the test accuracy has also converged to a stable value. Whilst the persistent homology dimension has mostly converged by this point, its mean value of 2.55 further converges to 2.37 by epoch 120. Mean of three seeds, $\pm\sigma$ shaded.

Experiment details: in this experiment we explore the significance of stopping training once 100% training accuracy has been achieved; does persistent homology dimension remain stable after this? We train using a learning rate of 10^{-2} for 120 epochs, pausing every 10 epochs to measure $\dim_{\text{PH}}$. We note that since we are computing $\dim_{\text{PH}}$ using $K = 1000$ iterations of weights (corresponding to around 2.5 epochs), the calculations are of limited granularity. To permit comparisons with the results of Birdal et al. (2021), we do not use BatchNorm. The results for these experiments are presented in Figure 4.

Discussion: the results of this experiment demonstrate that whilst $\dim_{\text{PH}}$ has mostly converged when 100% training accuracy is achieved, there is a period of continued convergence after this point. We further note the convergence of $\dim_{\text{PH}}$ to be both stable and monotonic. The experiments of Birdal et al. (2021) all define convergence as reaching 100% training accuracy. However, this experiment indicates that $\dim_{\text{PH}}$ continues to decrease by a further 0.18 beyond this point. This is non-negligible when contextualised with the results in Figure 3 (left), where the learning rate clusters differ in $\dim_{\text{PH}}$ by approximately 0.25. However, referring to Theorem 1 we see that the bound is provided for the supremum over elements within the trajectory, in this case the K preceding iterations. A portion of this continued descent in persistent homology dimension could be attributed to this, particularly given that the following evaluation does not occur until epoch 80. Further work could explore the convergence rate of persistent homology dimension at a range of learning rates, of particular interest would be if the relationships in Figure 3 hold after more extensive training.

7 Model-Wise Double Descent

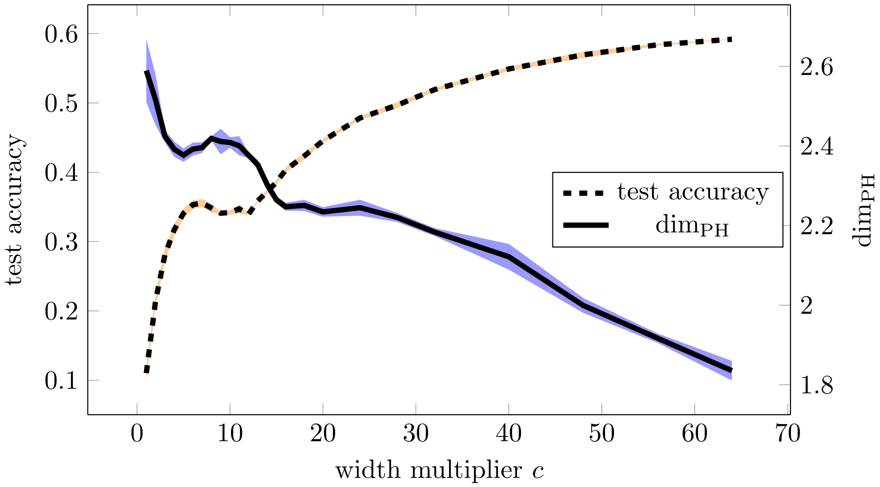

Figure 5: Model-wise double descent manifests in both test accuracy and persistent homology dimension $\dim_{\mathrm{PH}}$**.** Model-wise double descent refers to non-monotonicity with respect to model parameter count, which is evident in both test accuracy and persistent homology dimension for increasing channel width multiplier c. Mean of three seeds, $\pm\sigma$ shaded.

Experiment details: we now consider model-wise double descent; the phenomenon in which test accuracy is non-monotonic with respect to increasing model parameters. We closely follow the procedure of Nakkiran et al. (2019) (see their Figure 7), training on CIFAR-100 with BatchNorm. Label noise can be added to accentuate the double descent behaviour, but this was omitted from our experiments. We follow Nakkiran et al. (2019) in using a decaying learning rate schedule. We scale the channel width multiplier c to values in the range $[0, 64]$ with particularity high density up to 20 where double descent is present. We train for 500 epochs regardless of train accuracy convergence, but note that no model achieves 100% train accuracy on the harder CIFAR-100 dataset in this period. The results for these experiments are presented in Figure 5.

Discussion: these results demonstrate model-wise double descent in persistent homology dimension. The double descent peak occurs in the same range of width multipliers for both metrics, although persistent homology dimension plateaus for a number of width multiplier values following its peak. This is an interesting, and surprising result, suggesting that while persistent homology dimension does not measure generalisation, there is a connection with double descent, possibly in the behaviour of the training dynamics once the model has converged. We remark that double descent is a phenomenon in test accuracy, and not generalisation error, accordingly Figure 5 plots test accuracy (which is *not* $1 -$ generalisation due to 100% training accuracy not being achieved). It is therefore interesting that persistent homology dimension exhibits double descent despite the generalisation error monotonically increasing (as the train accuracy improves at lower c than test) in this region of width multipliers. We further comment that changing width multiplier c will lead to a different constants M in Theorem 1, encouraging further work to understand the connection to this bound.

8 Conclusion

This project has presented experimental evidence that persistent homology dimension does not measure generalisation. The two identified failures modes are large learning rates, and adversarial initialisation. These results lead us to hypothesise that persistent homology dimension is simply a measure of *convergence* that is additionally correlated with learning rate, where the experiments of Birdal et al. (2021) did not succeed to ablate these spurious correlations. The adversarial initialisation experiment is supportive of the convergence hypothesis, since both models have converged (and thus have low $\dim_{\mathrm{PH}}$ but only one has strong generalisation). We believe the result for double descent can additionally be understood in this framework, where the convergence behaviour of the models in the double descent 'peak' have some connection to $\dim_{\mathrm{PH}}$. Fully understanding the connection to double descent remains exciting future work.

With respect to learning rate, we believe the correlation with learning rate is the dominant effect in Figure 3, with generalisation simply correlating with learning rate at certain values. Since only the zeroth dimension persistent homology is considered, the lifetimes of interest all start at a scale of zero, and end as the components are connected, see Figure 2. We are therefore only interested in the distance between each point and its nearest neighbour (since once it connects a component will have died). It is therefore intuitive that the more spaced out the final K iterates W are, the more sensitive the lifetime sum will be to the number of samples n taken when computing W_n (since the points are so far apart adding more samples leads to a rapid decrease in lifetime sum). This sensitivity manifests as a high gradient m and therefore a low $\dim_{\mathrm{PH}}$ value. Using a large learning rate will lead to larger steps in the final K iterates, making this a possible explanation for the correlation between $\dim_{\mathrm{PH}}$ and learning rate. A formal proof of this intuition warrants further investigation.

References

Adams, H. (2021). DSCI 475: Topological Data Analysis.

Adams, H., Aminian, M., Farnell, E., Kirby, M., Mirth, J., Neville, R., Peterson, C., and Shonkwiler, C. (2020). A Fractal Dimension for Measures via Persistent Homology. In Baas, N. A., Carlsson, G. E., Quick, G., Szymik, M., and Thaule, M., editors, *Topological Data Analysis*, Abel Symposia, pages 1–31, Cham. Springer International Publishing.

Bartlett, P. L. and Mendelson, S. (2003). Rademacher and gaussian complexities: risk bounds and structural results. *The Journal of Machine Learning Research*, 3(null):463–482.

Birdal, T., Lou, A., Guibas, L. J., and Simsekli, U. (2021). Intrinsic Dimension, Persistent Homology and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 6776–6789. Curran Associates, Inc.

Chazal, F. and Michel, B. (2021). An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. *Frontiers in Artificial Intelligence*, 4.

Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2000). Topological persistence and simplification. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 454–463. ISSN: 0272-5428.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D., and Wilson, A. G. (2018). Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. arXiv:1802.10026 [cs, stat].

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer. Google-Books-ID: eBSgoAEACAAJ.

He, H., Huang, G., and Yuan, Y. (2019). Asymmetric Valleys: Beyond Sharp and Flat Local Minima. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Liu, S., Papailiopoulos, D., and Achlioptas, D. (2021). Bad Global Minima Exist and SGD Can Reach Them. arXiv:1906.02613 [cs, stat].

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2019). Deep Double Descent: Where Bigger Models and More Data Hurt. arXiv:1912.02292 [cs, stat].

Rieck, B., Togninalli, M., Bock, C., Moor, M., Horn, M., Gumbsch, T., and Borgwardt, K. (2019). Neural Persistence: A Complexity Measure for Deep Neural Networks Using Algebraic Topology. page 25 p. arXiv:1812.09764 [cs, math, stat].

Simsekli, U., Sener, O., Deligiannidis, G., and Erdogdu, M. A. (2020). Hausdorff Dimension, Heavy Tails, and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 5138–5151. Curran Associates, Inc.

Smith, S. L., Dherin, B., Barrett, D. G. T., and De, S. (2021). On the Origin of Implicit Regularization in Stochastic Gradient Descent. arXiv:2101.12176 [cs, stat].

Valle-Pérez, G. and Louis, A. A. (2020). Generalization bounds for deep learning. arXiv:2012.04115 [cs, stat].

Wikipedia (2023). Homology (mathematics). Page Version ID: 1131206780.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. arXiv:1611.03530 [cs].

Zomorodian, A. and Carlsson, G. (2004). Computing persistent homology. In *Proceedings of the twentieth annual symposium on Computational geometry*, SCG '04, pages 347–356, New York, NY, USA. Association for Computing Machinery.