Metric Space Magnitude and Generalisation in Neural Networks

Rayna Andreeva [1] **Katharina Limbeck** [2 3] **Bastian Rieck** [† 2 3] **Rik Sarkar** [† 1]

Abstract

Deep learning models have seen significant successes in numerous applications, but their inner workings remain elusive. The purpose of this work is to quantify the learning process of deep neural networks through the lens of a novel topological invariant called *magnitude*. Magnitude is an isometry invariant; its properties are an active area of research as it encodes many known invariants of a metric space. We use magnitude to study the internal representations of neural networks and propose a new method for determining their generalisation capabilities. Moreover, we theoretically connect magnitude dimension and the generalisation error, and demonstrate experimentally that the proposed framework can be a good indicator of the latter.

1. Introduction

Deep neural networks (DNNs) have become ubiquitous due to their remarkable performance in a range of tasks, including computer vision (Xie et al., 2017), natural language processing (Vaswani et al., 2017), and scientific discovery (Jumper et al., 2021). However, state-of-the-art DNNs often comprise millions to billions of parameters, making it impractical for human users to gain a precise understanding of the inner workings of these networks. One crucial yet unanswered question is to understand the generalisation of neural networks, i.e. their ability to perform well on unseen data. In recent literature, various notions of neural networks' intrinsic dimensions have been proposed (Birdal et al., 2021; Dupuis et al., 2023; Simsekli et al., 2020), which demonstrate that the most important ingredient for generalisation is effective capacity rather than the number of parameters.

In this work, we propose a novel measure of the intrinsic dimension of neural networks, based on a novel topological invariant called magnitude. Unlike previous approaches, magnitude benefits from a simple interpretation, and potentially better computational complexity. This choice is also theoretically justified. Magnitude is an isometry invariant of a metric space and its properties are an active area of mathematical research due to the fact that it encodes many important invariants from geometric measure theory and integral geometry (Leinster, 2013). Further, magnitude has roots in theoretical ecology and it can capture the effective number of species in an ecosystem. In a broader context, it has been shown that magnitude can thus encode the effective number of distinct points in a space (Leinster, 2013). We further build on this idea and propose a novel method for evaluating the generalisation error using the concept of an effective number of models, which acts as an effective capacity.

Although there has been significant theoretical research on magnitude, a considerable number of its characteristics are yet to be discovered, and several unanswered questions persist, particularly regarding its practical applications in machine learning. While recent studies (Adamer et al., 2021; Bunch et al., 2021) have started to establish links between magnitude vectors and machine learning applications, this area of research is still in its very early stages.

Recently, it has been shown that a concept derived from magnitude, known as the magnitude dimension, is the same as the Minkowski dimension (Meckes, 2015) under certain conditions. As a result, we can theoretically show that the generalisation error of the trajectories of a training algorithm is intrinsically linked to the magnitude dimension of the so-defined metric space. This enables us to generate novel insights about the learning process of neural networks. Further, our contribution is the first work exploring the magnitude dimension, which has solid theoretical foundations, as a notion of the intrinsic dimension of neural networks.

Contributions. Our work is the first to introduce the mathematical concept of magnitude into theoretical deep learning and the study of neural network generalisation; we believe that this is an exciting novel contribution to the nascent field of topological machine learning (Hajij et al., 2022; Hensel et al., 2021). After establishing a theoretical bound, we explore the change in the magnitude function across multiple

[1]School of Informatics, University of Edinburgh [2]Helmholtz Munich [3]Technical University of Munich. Correspondence to: Rayna Andreeva <r.andreeva@sms.ed.ac.uk>.

Proceedings of the 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML) at the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. 2023. *Copyright 2023 by the author(s).*

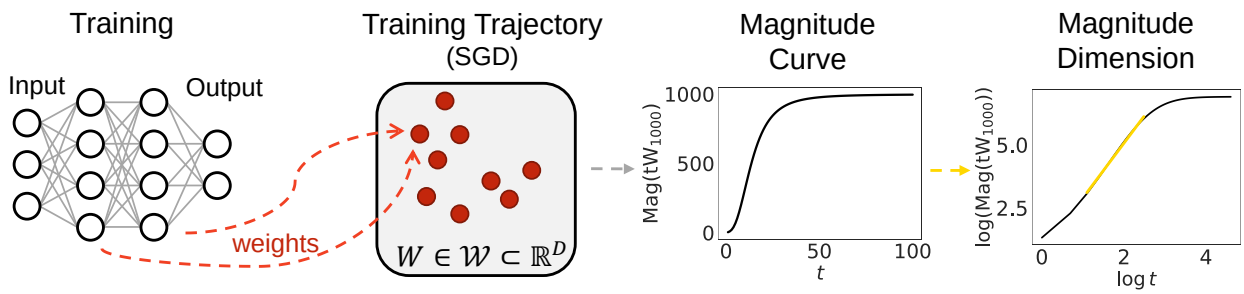


Figure 1. **Overview of the procedure.** We first train a neural network and monitor the training trajectory. At each 1000 iterations of the trajectory, we collect the training weights $W \in \mathbb{R}^d$ into a point cloud. Then, we compute the magnitude curve of the point cloud at selected scales t, create the log-log plot and estimate the magnitude dimension based on it.

experimental settings, and compare the value of magnitude at multiple scales to the test accuracy. Further, we demonstrate experimentally that there is a link between magnitude and the test accuracy. By making a novel connection with the persistent homology dimension, we then compare our proposed measure with the intrinsic dimension introduced by Adams et al. (2020) and demonstrate that ours benefits from a better computational complexity and interpretability. In short, our contributions are as follows:

- We propose a novel method for evaluating the generalisation of neural networks based on magnitude and the effective number of models, which allows us to monitor performance without a validation set.
- We prove a new upper bound for the generalisation error, linking the generalisation error to a magnitude-based characteristic of the training trajectories.
- We empirically show that the evolution of these measures throughout the training process correlates with the accuracy of the test set.
- We prove that all notions of previously proposed intrinsic dimensions are the same as the magnitude dimension, and we verify this result empirically.

2. Related Work

We briefly review the literature related to generalisation in neural networks, intrinsic dimension, and magnitude.

Generalisation Bounds and Intrinsic Dimension. Several works explore intrinsic dimension for capturing the generalisation capabilities of neural networks. Simsekli et al. (2020) demonstrated that the fractal dimension of a hypothesis class is associated with the generalisation error, which is further linked to the heavy-tailed behavior of the trajectory of networks (Hodgkinson & Mahoney, 2021; Mahoney & Martin, 2019; Simsekli et al., 2019). However, many assumptions were required for the bound to be computed in practice. A more recent work relaxed some of the assumptions, and developed the notion of the persistent homology dimension (Adams et al., 2020), $\dim_{\mathrm{PH}}$. The authors in Birdal et al. (2021) were the first to offer a theoretical justification for using topological invariants for the analysis of deep neural networks. Another work (Magai & Ayzenberg, 2022) investigated $\dim_{\mathrm{PH}}$ at different depths and layers of the network and observed its evolution. In Dupuis et al. (2023), the authors developed a data-driven dimension and compared it with the $\dim_{\mathrm{PH}}$ of Birdal et al. (2021). They have demonstrated stronger correlation with the generalisation error than previously shown and managed to relax some of the restrictive assumptions.

Magnitude and its Applications in Machine Learning. Magnitude was first proposed in Solow & Polasky (1994) for measuring biodiversity, albeit without any reference to its mathematical properties. It was only approximately twenty years later when Leinster (2013) formalised its mathematical properties using the language of category theory. Further, magnitude has been realised as the Euler characteristic in magnitude homology (Leinster & Shulman, 2021). While magnitude has theoretical foundations, its applications to machine learning are scarce. Recently, there has been renewed interest in introducing magnitude into the machine learning community. The first work to develop the concept of magnitude in the context of machine learning demonstrated that the individual summands of magnitude, known as magnitude weights, can be used as an efficient boundary detector (Bunch et al., 2021). Further, it has been used for working in the space of images and it has demonstrated its usefulness as an effective edge detector (Adamer et al., 2021). However, our contribution constitutes the first direct application of magnitude to deep learning.

Using Topology to Characterise Neural Networks. Earlier research has established a connection between neural network training and topological invariants (Fernández et al., 2021), using topological complexity as proxy for generalisation performance, for instance (Rieck et al., 2019). However, these studies focused solely on analyzing the trained network after completing the training process (Fernández et al.,

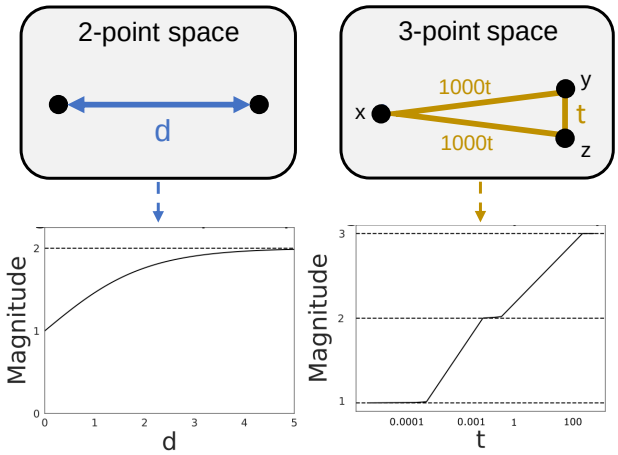


Figure 2. **Magnitude of the 2- and 3-point space.** On the left, we see the 2-point space, where the distance between the points is d. On the right we see the 3-point space for an isosceles triangle with distance t between y and z, and $1000t$ between x and y and x and z. Below each space, we see the respective magnitude function.

2021), potentially missing critical aspects of the training dynamics (Birdal et al., 2021). By contrast, we propose the use of another topological invariant—magnitude—which affords more interpretability than previous approaches. Moreover, we compute magnitude on the training trajectories instead of on the trained network, offering crucial topological insights into training dynamics.

3. Background

We start by defining magnitude and the magnitude dimension, and then proceed to define intrinsic dimensions.

3.1. Magnitude

While the magnitude of metric spaces is a general concept (Leinster, 2013), we restrict our focus to subsets of $\mathbb{R}^n$, where magnitude is known to exist (Meckes, 2013).

Definition 3.1. Let (X, d) be a finite metric space with distance metric d. Then, the similarity matrix of X is defined as $\zeta_{ij} = e^{-d_{ij}}$ for $1 \leq i, j \leq n$, where n denotes the cardinality of X.

Definition 3.2. Let X be a metric space with similarity matrix ζ_{ij}. If ζ_{ij} is invertible, magnitude is defined as

$$\mathrm{Mag}(X) = \sum_{ij} (\zeta^{-1})_{ij}. \tag{1}$$

When X is a finite subset of $\mathbb{R}^n$, then ζ_{ij} is a symmetric positive definite matrix as proven in Leinster (2013) (Theorem 2.5.3). Then, $(\zeta^{-1})_{ij}$ exists, and hence magnitude exists as well. Magnitude is best illustrated when considering a few sample spaces with a small number of points.

Example 3.3. Let X denote the metric space with a single point a. Then, ζ_X is a 1×1 matrix with $\zeta_X^{-1} = 1$ and using the formula for magnitude, we get $\mathrm{Mag}_X = 1$.

Example 3.4. A more illustrative example is given by the space of two points. Let $X = \{a, b\}$ be a finite metric space where $d_X(a, b) = d$. Then

$$\zeta_X = \begin{bmatrix} 1 & e^{-d} \\ e^{-d} & 1 \end{bmatrix}, \tag{2}$$

so that

$$\zeta_X^{-1} = \frac{1}{1 - e^{-2d}} \begin{bmatrix} 1 & -e^{-d} \\ -e^{-d} & 1 \end{bmatrix}, \tag{3}$$

and therefore

$$\mathrm{Mag}(X) = \frac{2 - 2e^{-d}}{1 - e^{-2d}} = \frac{2}{1 + e^{-d}}. \tag{4}$$

This example is also illustrated in Figure 2. Using Eq. (4), if d is very small, i.e. $d \to 0$, $\mathrm{Mag}(X) \to 1$. Similarly, when $d \to \infty$, $\mathrm{Mag}(X) \to 2$. In practice, as it can be seen from Figure 2, $\mathrm{Mag}(X) = 2$ for a value of d as small as 5.

More information about a metric space can be obtained by looking at its rescaled counterparts. The resulting representation is richer, and it can be summarised compactly in the magnitude function. For this purpose, the scale parameter t is introduced. By varying t, we obtain a scaled version of the metric space, which permits us to answer what the number of *effective points* of a metric space is.

The Magnitude Function. Magnitude assigns to each metric space not just a scalar number, but a function. This works as follows: we introduce a scale parameter t, and for each value of t, we consider the space where the distances between points are scaled by t. More concretely, for $t > 0$, the notation tX means X scaled by a factor of t. Computing the magnitude for each value of t then gives the magnitude function. More formally, we have the following definition:

Definition 3.5. Let (X, d) be a finite metric space. We define (tX, d_t) to be the metric space with the same points as X and the metric $d_t(x, y) = td(x, y)$.

The intuition here is that we are looking at the same space but through different scales, for example, changing the distances from metres to centimeters. Then the definition of the magnitude function follows naturally.

Definition 3.6. The magnitude function of a finite metric space (X, d) is the function $t \mapsto \mathrm{Mag}(tX)$, which is defined for all $t \in (0, \infty)$.

Example 3.7. Consider the magnitude function of the 3-point space in Figure 2. In this example, the points x, y

and z form an isosceles triangle. When $t = 0.0001$, all the three points are very close to each other and almost indistinguishable. Hence, we say that the space has 1 effective point. In contrast, when $t = 0.01$, the distance between the two points on the right y and z is very small, and x is quite far. Therefore, we say that the space looks like two points. Finally, when t is large, all the three points are far away from each other, and the value of magnitude is 3, which is also the cardinality of the space.

3.2. Intrinsic Dimension

There are various notions that can be used to measure the intrinsic dimension of a space. In this work, we will focus on three such notions: the upper-box dimension (Minkowski), the magnitude dimension and the persistent homology dimension. The box dimension is based on covering numbers and can be linked to generalization via Simsekli et al. (2020), whereas the magnitude dimension is built upon the concepts we defined earlier.

Definition 3.8 (Minkowski dimension)**.** For a bounded metric space X, let $N_\delta(X)$ denote the maximal number of disjoint closed δ-balls with centers in X. The upper box/Minkowski dimension is defined as

$$\dim_{\text{Mink}} X = \lim_{\delta \to 0} \sup \frac{\log(N_\delta(X))}{\log(\frac{1}{\delta})}. \tag{5}$$

There is a subtle point to be made here. In general, the Minkowski and Hausdorff dimensions do not coincide and are not equivalent. However, in Simsekli et al. (2020) the authors provide conditions under which the Hausdorff dimension of the space we are interested in coincides with the Minkowski dimension. In fact, for many fractal-like sets, these two notions of dimensions are equal to each other; see Mattila (1999, Chapter 5).

Definition 3.9 (Magnitude dimension)**.** When

$$\dim_{\text{Mag}} X = \lim_{t \to \infty} \frac{\log(\text{Mag}(tX)))}{\log t} \tag{6}$$

exists, we define this to be the magnitude dimension of X (Meckes, 2015).

The magnitude dimension can be approximately interpreted as the rate of change of the magnitude function for a suitable interval of values for t. We can introduce another notion of the fractal dimension, known as the persistent homology dimension ($\dim_{\text{PH}}$) (Adams et al., 2020).

Definition 3.10. The persistent homology dimension of a bounded metric space (X, d), denoted by $\dim^0_{\text{PH}}$, is defined as

$$\inf\{\alpha > 0, \exists C > 0, \forall W \subset X \text{finite}, E_\alpha < C\}, \tag{7}$$

where E_α is the α-lifetime sum, defined as

$$E_\alpha(W) := \sum_{(b,d) \in PH^0(Rips(W))} (d - b)^\alpha,$$

and b and d are the birth and death values respectively for the persistent homology of degree 0 (PH^0). It measures all connected components in the Vietoris-Rips filtration of W, denoted by $Rips(W)$.

The definition is rather technical and it is not crucial for the work here, for more details please refer to Adams et al. (2020); Edelsbrunner & Harer (2022); Mémoli & Singhal (2019). We note that it can be defined for persistent homology of any degree, but for the purpose of this paper, we are only interested in degree 0. Therefore, for ease of notation, we will omit the 0 and denote $\dim^0_{\text{PH}}$ by $\dim_{\text{PH}}$, i.e. $\dim_{\text{PH}} := \dim^0_{\text{PH}}$.

4. Theoretical Results

We first elucidate connections between different notions of intrinsic dimension before proving connections to the generalisation error.

4.1. Connection between Notions of Intrinsic Dimension

After we have introduced three different notions of intrinsic dimensions, we demonstrate that they are in fact the same under some mild assumptions. Our novel contributions is proving that $\dim_{\text{Mag}} X$ and $\dim^0_{\text{PH}} X$ are equal. The result is formalised in the following theorem, which assumes that all notions of dimension exist.

Theorem 4.1. *Let $X \subset \mathbb{R}^n$ be a compact set and either $\dim_{\text{Mag}} X$ or $\dim_{\text{Mink}} X$ exist. Then*

$$\dim_{\text{Mag}} X = \dim^0_{\text{PH}} X \tag{8}$$

Proof. Since X is compact and either $\dim_{\text{Mag}} X$ or $\dim_{\text{Mink}} X$ exist, from Corollary 7.4, (Meckes, 2015) it follows that both $\dim_{\text{Mag}} X$ and $\dim_{\text{Mink}} X$ exist and $\dim_{\text{Mag}} X = \dim_{\text{Mink}} X$. Since X is compact, from the Heine-Borel Theorem, X is both closed and bounded, and therefore from a result in Kozma et al. (2006); Schweinhart (2021), we have that $\dim_{\text{Mink}} X = \dim^0_{\text{PH}} X$. Hence, $\dim_{\text{Mag}} X = \dim_{\text{Mink}} X = \dim^0_{\text{PH}} X$, which implies that $\dim_{\text{Mag}} X = \dim^0_{\text{PH}} X$.

□

4.2. Connection to the Generalisation Error

After having established equality between the various notions of dimensions, we proceed to formalise the required

language of machine learning theory, culminating in a novel generalisation result.

For the beginning of this section, we follow the notation from Shalev-Shwartz & Ben-David (2014). We briefly recall some standard definitions to make our paper self-contained. In a standard statistical learning setting, $\mathcal{X}$ denotes the set of features and $\mathcal{Y}$ the set of labels. Together, the cross product $\mathcal{X} \times \mathcal{Y}$ represents the space of data $\mathcal{Z}$. The learner has access to a sequence of data of m samples, called the training data, denoted by $S = ((x_1, y_1), ..., (x_m, y_m))$ in $\mathcal{X} \times \mathcal{Y} = \mathcal{Z}$. We assume that the training set S is generated by some unknown probability distribution over $\mathcal{X}$, which we denote by $\mathcal{D}$, and that each of the samples $\{x_i, y_i\}$ are independent and identically distributed (i.i.d) samples from $\mathcal{D}$. We will focus on a restricted search space for finding a set of predictors, which we call a hypothesis class, $\mathcal{H}$. Each element, called a hypothesis, $h \in \mathcal{H}$, is a function from $\mathcal{X}$ to $\mathcal{Y}$. An optimal hypothesis, or parameter vector $h \in \mathcal{H}$ is selected by computing a quantity called a loss function, defined by $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$. The empirical error is then $L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h, z)$ for $z \in \mathcal{Z}$ and the true error or risk is $L_{\mathcal{D}}(h) = \mathbb{E}_z[\ell(h, z)]$ for $z \in \mathcal{Z}$. The generalisation error is defined as the difference between the true error and the empirical risk, or $|L_S(h) - L_{\mathcal{D}}(h)|$.

In the case of neural networks, the hypothesis class is $\mathcal{W} \subset \mathbb{R}^d$, and each $w \in \mathcal{W}$ is a parameter vector. Given a training algorithm $\mathcal{A}$, we want to study the set of all hypothesis classes returned from the optimisation procedure $\mathcal{A}$ for a given training data S. We call this the *optimisation trajectories*, and denote them by $\mathcal{W}$. To access the iterates at every step of the process, we denote by w_i an element of $\mathcal{W}$ at iteration i. More concretely, $\mathcal{W} := \{w \in \mathbb{R}^d : \exists i \in [0, I], w = [\mathcal{A}(S)_i]\}$, where I is the number of training iterations. In other words, when we fix i, we are interested in the weights at iteration i, returned by the optimisation algorithm $\mathcal{A}$.

For the following result, there is a more technical condition required from Birdal et al. (2021, Assumption H1), which can be found in the Appendix. We denote this assumption as H1. We require the existence of the constant $M > 0$, which quantifies how dependent the set $\mathcal{W}_S$ is on the training sample S. Now that we have all the required definitions, we can proceed with the novel result.

Theorem 4.2. *Let $\mathcal{W} \in \mathbb{R}^d$ be a compact set. Under the assumption that $H1$ holds, ℓ is bounded by a constant C and K-Lipschitz continuous in w. For n sufficiently large, we then have the following bound:*

$$\sup_{w \in \mathcal{W}} |L_S(w) - L_{\mathcal{D}}(w)| \leq 2C\sqrt{\frac{[\dim_{\mathrm{Mag}}\mathcal{W} + 1]\log^2(nK^2)}{n} + \frac{\log(7M/\gamma)}{n}} \tag{9}$$

with probability $1 - \gamma$ over $S \sim \mathcal{D}^{\otimes n}$, where M is the constant from Assumption H1.

Proof. Since $\mathcal{W}$ is bounded, we have $\dim^0_{\mathrm{PH}}\mathcal{W} = \dim_{\mathrm{Mag}}\mathcal{W}$. Therefore, the result follows from substituting $\dim^0_{\mathrm{PH}}\mathcal{W}$ with $\dim_{\mathrm{Mag}}\mathcal{W}$ in Proposition 1 (Birdal et al., 2021). □

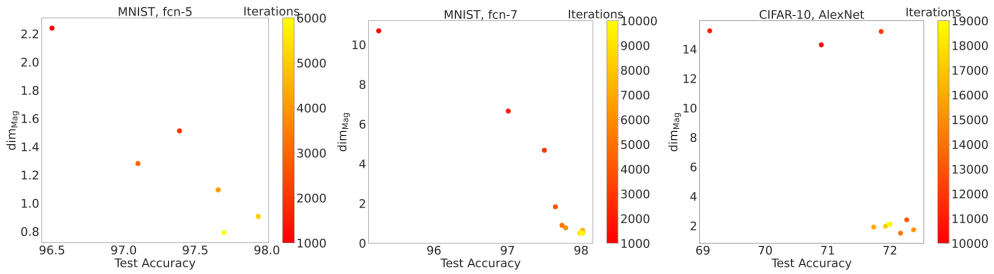
5. Methods

In contrast with previous work on the intrinsic dimension, we propose to estimate the fractal dimension using the magnitude. As we have seen, the concept of magnitude dimension coincides with the Minkowski dimension and the persistent homology dimension. This theoretical connection enables us to confidently explore the concept of magnitude in the context of neural networks. We do this as follows: at selected points on the weight trajectory $\mathcal{W}$, we subsample a number of models W, which can be interpreted as a point cloud, where each point is a model in the model space. This space has a very high dimension equal to the number of parameters in the network. We compute the magnitude and the magnitude dimension of each such point cloud. We then investigate the connection between these quantities and the test accuracy.

In light of our novel result in Theorem 4.2, linking the intrinsic dimension and the generalisation bound, it is natural to ask if the magnitude function can also be used to explore the learning process of neural networks. Since the magnitude dimension can be roughly interpreted as the rate of change of the magnitude function, and therefore, if there is a link between the magnitude dimension and the generalisation error, then there should also be a link between the values of the magnitude function at different scales and the generalisation error. What we want to study is the change of the space of model trajectories as the learning process advances. In other words, does the space look like a bigger number of distinct points or does it resemble fewer points as the training progresses? Translating this idea to the language of magnitude, is the *effective number of models* increasing or decreasing when performing more iterations of the learning algorithm?

Since magnitude is a function, we would like to take a cross-sectional slice of the magnitude curve and examine the magnitude values for a particular choice of the scale parameter t. Therefore, we formulate the definition of the effective number of models for a fixed t.

Definition 5.1. We define the **effective number of models** as the value of $\mathrm{Mag}(t_i\mathcal{W})$ at scale t_i.



(a) (b) (c)

Figure 3. **The magnitude dimension correlates with the test accuracy.** In plots (a-c), we see the magnitude dimension plotted against the test accuracy over a varying number of iterations, which are depicted in different colours, from red to yellow. We note that there is correlation between the magnitude dimension and the test accuracy across both MNIST and CIFAR-10, and across different architectures (FCN-5, FCN-7, AlexNet), which is stronger for MNIST than for CIFAR-10.

Algorithm 1 Estimation of $\dim_{\text{Mag}}\mathcal{W}$

Input: The set of the training trajectories $\mathcal{W} = \{w_i\}^n$ of size n, scale parameter $t : [0, t_k]$, interval $[t_i, t_j], i < j < k$
Output: $\dim_{\text{Mag}}$
for $r = 1$ **to** k **do**
  $\text{Mag}(t_r\mathcal{W}) \leftarrow \mathcal{W}$
end for
$m, b \leftarrow \text{fitline}(log(\text{Mag}(\mathcal{W}_n)[t_i : t_j]), log([t_i : t_j]))$
$\dim_{\text{Mag}}\mathcal{W} \leftarrow m$

5.1. Analyzing Deep Neural Network Dynamics via the Magnitude Dimension

Estimating the magnitude dimension is not a straightforward task, as the limit from Equation 6 needs to be approximated by finding the longest straight part of the magnitude function, which cannot be computed automatically, but needs to be done manually. Here we provide the algorithm for computation of $\dim_{\text{Mag}}\mathcal{W}$ from a finite sample W, approximating an infinite process. We follow the procedure similar to the estimations of the magnitude dimension (O'Malley et al., 2023; Willerton, 2009). The details are described in Algorithm 1. After choosing a suitable representative interval $[t_i, t_j]$, the log-log plot of magnitude versus t is generated, and the slope m of the line and the intercept b are computed at the selected interval $[t_i, t_j]$. Then, m is taken to be the estimate of $\dim_{\text{Mag}}\mathcal{W}$. This procedure can be essentially seen as computing the limit based on the slope, which is an application of l'Hôpital's rule.

6. Experimental Results

The first goal of this section is to verify our main claim, namely that our estimate of the magnitude dimension is capable of measuring generalisation. The second goal is to establish a connection between magnitude itself and the test accuracy. The third goal is to empirically compare the two very close measures of the fractal dimension, namely $\dim_{\text{Mag}}\mathcal{W}$ and $\dim_{\text{PH}}\mathcal{W}$. The fourth goal is to attempt to explain what our results mean for generalisation in neural networks in general, arriving at novel insights. In order to show this, we use our procedure on a number of different neural network architectures, training settings and datasets. In particular, we train a 5-layer (fcn-5) and 7-layer (fcn-7) on MNIST and AlexNet (Krizhevsky et al., 2017) on CIFAR10, over a different range of learning rates and batch size of 100. We consider a sliding window of 1000 training iterations. We then estimate $\dim_{\text{Mag}}\mathcal{W}$ of each window, following the steps in Algorithm 1.

6.1. Exploring the learning process

We assess the main claim of the paper, which links the magnitude dimension with the generalisation error. Figure 3 reveals multiple findings. First, we observe that there is a correlation between the magnitude dimension and the test accuracy—the lower the magnitude dimension, the higher the test accuracy. This result agrees with what has previously been observed in Birdal et al. (2021); Simsekli et al. (2020), albeit from the perspective of *magnitude*, an invariant that is arguably simpler to compute and more interpretable in practice than persistent homology. Second, this holds across both MNIST and CIFAR10, and also across different architectures, which shows that even though the parameters differ considerably, the intrinsic dimension of different datasets can still be similar. To achieve good generalisation performance, it is important to keep the dimension as small as possible, without losing important representational features by collapsing them onto the same dimension.

6.2. Analysing and visualising network trajectories

Next, we want to investigate the effect on the magnitude function as the network is trained for more iterations. For

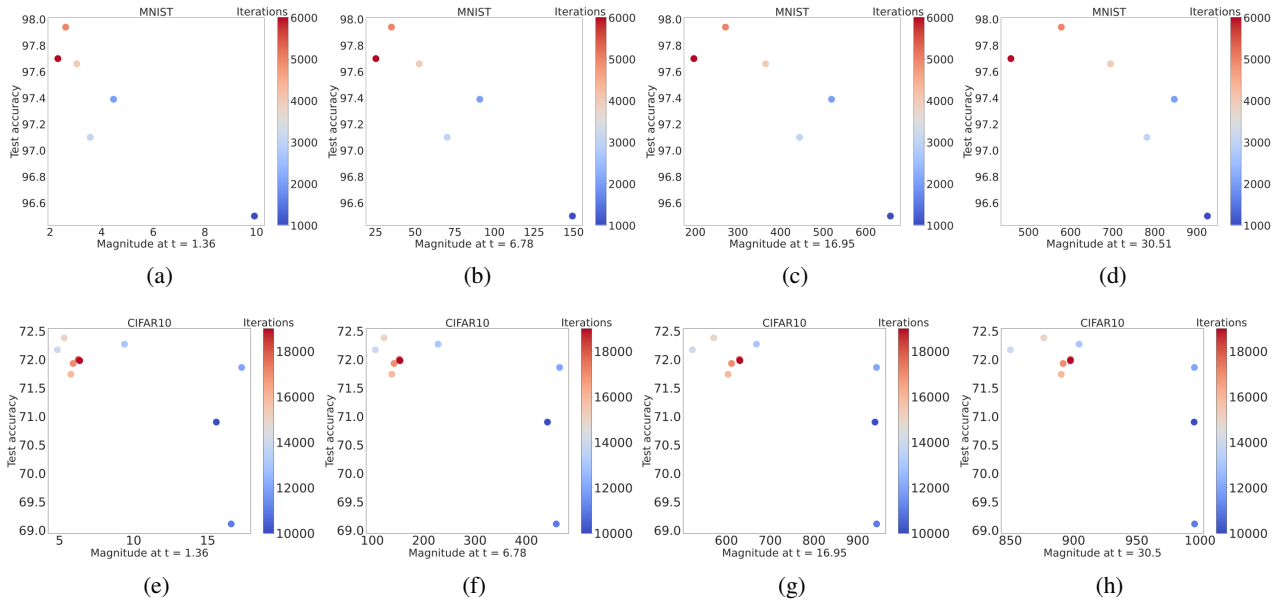


Figure 4. **The effective number of models correlates with the test accuracy.** Here we see the cross-sectional evaluation of the magnitude curve at different values of t. Each point represents the model trajectory over 1000 iterations, over which the magnitude is computed, as well as the test accuracy at the last model in the sliding window. The first row (plots (a-d)) shows the magnitude for MNIST. The second row (plots (e-h)) shows the plots for CIFAR10. We note that across all scales, there is similar pattern of correlation between the test accuracy and the magnitude values.

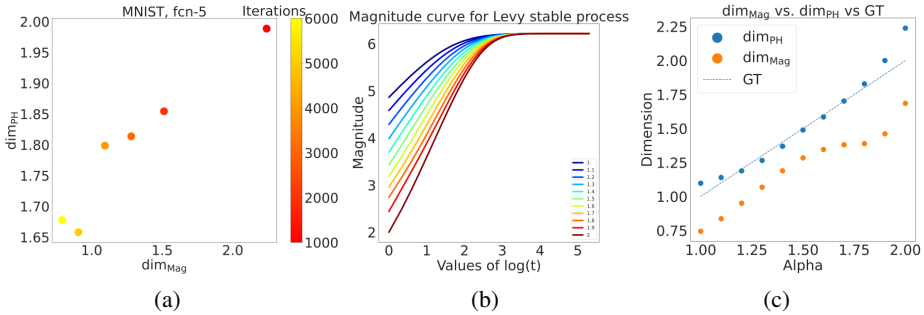


Figure 5. **The magnitude dimension, persistent homology dimension and ground truth for an α-Levy stable process.** In plot (a) we compare the $\dim_{\mathrm{Mag}}\mathcal{W}$ and $\dim_{\mathrm{PH}}\mathcal{W}$ against the number of iterations. There is strong correlation of 0.96, with statistically significant p-value ($p < 0.05$). In plot (b) we see the magnitude curves for different values of α. In plot (c) we see the magnitude curves of the α-stable Levy process for multiple values of α.

this purpose, we select four values of t across the range $[0, 40]$, $t \in \{1.36, 6.78, 16.95, 30.51\}$ and we compute the effective number of models $\mathrm{Mag}(t\mathcal{W})$ to give us a glimpse into the space of training trajectories. We further fix the learning rate and vary the number of iterations. The experiments were performed with a learning rate of 0.1 for illustrative purpose. However, we note that this pattern holds for multiple learning rates. In Figure 4 we can see the resulting magnitude value at each of the selected scales of t, and the colour depicts the number of iterations. We note that there is a concentration of red points in the upper left corner, indicating that higher test accuracy corresponds to a lower value of magnitude. Similarly, for the blue points, the higher value of magnitude is linked to worse test set performance. This pattern holds across both MNIST and CIFAR10. Moreover, the lower the test accuracy, the higher the magnitude value, implying that models with lower magnitude values generalise better. This result intuitively makes sense. When the magnitude value is small, the space looks like a smaller number of points. With the increase of t, the magnitude converges to the cardinality of the set. The effective number of models in (a) equals approximately 2 for high test accuracy, which is an interesting phenomenon. This means that out of the 1000 models, for $t = 1.38$, the space looks like

2 models, indicating that there are 2 large clusters formed by the training trajectories. Surprisingly, this is not the case for network trained across 1000 iterations, with the lowest test accuracy. In that case, the space of models looks like 10 points and is more scattered.

Note that the model with high test accuracy has smaller magnitude even at larger scales, suggesting that the models exhibit some sort of clustering behavior. In the magnitude terminology, the effective number of distinct models is smaller when the network generalises better; a "good optimisation trajectory" has a lower magnitude than a "bad optimisation trajectory"; the learning algorithm is implicitly optimising the magnitude, by clustering the models into a small number of clusters. Throughout our experiments we thus observed strong correlation between magnitude itself and the test accuracy, which persists across the different scales, learning rates and datasets. Therefore, the main insight from the results in Figure 4 is that a small effective number of models is good for generalisation and it indicates that there is a clustering pattern.

6.3. Similarities between $\dim_{\mathrm{Mag}}\mathcal{W}$ and $\dim_{\mathrm{PH}}\mathcal{W}$

In Figure 5(a), we demonstrate that there is a strong correlation between $\dim_{\mathrm{Mag}}\mathcal{W}$ and $\dim_{\mathrm{PH}}\mathcal{W}$ on MNIST. Although there is a good correspondence between the two, $\dim_{\mathrm{PH}}\mathcal{W}$ is more difficult to interpret. In order to give some interpretation, the authors had to take a step back and produce the persistent diagrams which were used to calculate the specified dimension. However, due to the fact that they have to sample the space to estimate the dimension, these calculations involve a big number of different diagrams, hence linking $\dim_{\mathrm{PH}}\mathcal{W}$ with the original trajectory is not so straightforward. On the other hand, there is a more clear connection between magnitude and the magnitude dimension and the appearance of clusters of weight parameters.

Ablation study. In Figure 5 we see the results from an ablation study. We simulate data from a process similar to the weight trajectories with known ground truth, using a d-dimensional α-stable Levy process (Seshadri & West, 1982), where we take $d = 10$, and compare the values of the magnitude dimension to the fractal dimension. Further, we compare it with the persistent homology dimension, which as we have proven in Theorem 4.1, is the same as the magnitude dimension. The purpose of this study is to empirically evaluate this theoretical result. As seen in Figure 5(b), our dimension highly correlates with the ground truth. It seems that it is consistently lower than the true dimension by approximately 0.3. Nonetheless, the high statistically significant correlation indicates that the magnitude dimension is indeed very close to the fractal dimension for a process which resembles the iterations of SGD. This gives us high confidence that the magnitude dimension measures exactly what it is supposed to measure.

7. Conclusion

Our work provides the first connection between magnitude and the generalisation error. We proved a theoretical result linking the magnitude dimension of the optimisation trajectories and the test accuracy. We verified our results experimentally by exploring the evolution of magnitude across multiple experimental settings and datasets. We have further expanded our understanding about the clustering property of the weight trajectory: through both theoretical and empirical results, we demonstrated that models with better generalisation error tend to cluster more, compared to models with worse test performance, which are of higher magnitude and more spread out. This phenomenon has been previously described (Birdal et al., 2021; Brüel-Gabrielsson et al., 2019), and in this work our observations supplement it. In future work, we will improve on the estimation of the magnitude dimension and investigate the use of magnitude as a regulariser. Moreover, by using magnitude itself, we demonstrated that we can monitor the performance of the neural network without a validation set. In future work we can explore magnitude as an early stopping criteria, similar to Rieck et al. (2019). Furthermore, from our empirical ablation study, the magnitude dimension seems to be consistently lower than the ground truth, which will be investigated further. In particular, we want to consider to what extent it is possible to improve on the generalisation bound in Theorem 4.2.

Acknowledgements

The authors wish to thank Alexandros Keros and Emily Roff for helpful comments and insightful discussions that helped improve the manuscript. RA is supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics and the International Helmholtz-Edinburgh Research School for Epigenetics (EpiCrossBorders). KL is supported by Helmholtz Association under the joint research school "Munich School for Data Science - MUDS."

References

Adamer, M. F., De Brouwer, E., O'Bray, L., and Rieck, B. The magnitude vector of images. *arXiv preprint arXiv:2110.15188*, 2021.

Adams, H., Aminian, M., Farnell, E., Kirby, M., Mirth, J., Neville, R., Peterson, C., and Shonkwiler, C. A fractal dimension for measures via persistent homology. In *Topological Data Analysis: The Abel Symposium 2018*, pp.

1–31. Springer, 2020.

Birdal, T., Lou, A., Guibas, L. J., and Simsekli, U. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Bradley, R. C. On the ψ-mixing condition for stationary random sequences. *Transactions of the american mathematical society*, 276(1):55–66, 1983.

Brüel-Gabrielsson, R., Nelson, B. J., Dwaraknath, A., Skraba, P., Guibas, L. J., and Carlsson, G. A topology layer for machine learning. *arXiv preprint arXiv:1905.12200*, 2019.

Bunch, E., Kline, J., Dickinson, D., Bhat, S., and Fung, G. Weighting vectors for machine learning: numerical harmonic analysis applied to boundary detection. *arXiv preprint arXiv:2106.00827*, 2021.

Dupuis, B., Deligiannidis, G., and Şimşekli, U. Generalization bounds with data-dependent fractal dimensions. *arXiv preprint arXiv:2302.02766*, 2023.

Edelsbrunner, H. and Harer, J. L. *Computational topology: an introduction*. American Mathematical Society, 2022.

Fernández, D. P., Gutiérrez-Fandiño, A., Armengol-Estapé, J., and Villegas, M. Determining structural properties of artificial neural networks using algebraic topology. *arXiv preprint arXiv:2101.07752*, 2021.

Hajij, M., Zamzmi, G., Papamarkou, T., Miolane, N., Guzmán-Sáenz, A., Natesan Ramamurthy, K., Birdal, T., Dey, T. K., Mukherjee, S., Samaga, S. N., Livesay, N., Walters, R., Rosen, P., and Schaub, M. T. Topological deep learning: Going beyond graph data. *arXiv e-prints*, art. arXiv:2206.00606, 2022. doi: 10.48550/arXiv.2206.00606.

Hensel, F., Moor, M., and Rieck, B. A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, 4:681108, 2021.

Hodgkinson, L. and Mahoney, M. Multiplicative noise and heavy tails in stochastic optimization. In *International Conference on Machine Learning*, pp. 4262–4274. PMLR, 2021.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kozma, G., Lotker, Z., and Stupp, G. The minimal spanning tree and the upper box dimension. *Proceedings of the American Mathematical Society*, 134(4):1183–1187, 2006.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Leinster, T. The magnitude of metric spaces. *Documenta Mathematica*, 18:857–905, 2013.

Leinster, T. and Shulman, M. Magnitude homology of enriched categories and metric spaces. *Algebraic & Geometric Topology*, 21(5):2175–2221, 2021.

Magai, G. and Ayzenberg, A. Topology and geometry of data manifold in deep learning. *arXiv preprint arXiv:2204.08624*, 2022.

Mahoney, M. and Martin, C. Traditional and heavy tailed self regularization in neural network models. In *International Conference on Machine Learning*, pp. 4284–4293. PMLR, 2019.

Mattila, P. *Geometry of sets and measures in Euclidean spaces: fractals and rectifiability*. Number 44. Cambridge university press, 1999.

Meckes, M. W. Positive definite metric spaces. *Positivity*, 17(3):733–757, 2013.

Meckes, M. W. Magnitude, diversity, capacities, and dimensions of metric spaces. *Potential Analysis*, 42(2):549–572, 2015.

Mémoli, F. and Singhal, K. A primer on persistent homology of finite metric spaces. *Bulletin of mathematical biology*, 81:2074–2116, 2019.

O'Malley, M., Kalisnik, S., and Otter, N. Alpha magnitude. *Journal of Pure and Applied Algebra*, pp. 107396, 2023.

Rieck, B., Togninalli, M., Bock, C., Moor, M., Horn, M., Gumbsch, T., and Borgwardt, K. Neural persistence: A complexity measure for deep neural networks using algebraic topology. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=ByxkijC5FQ.

Schweinhart, B. Persistent homology and the upper box dimension. *Discrete & Computational Geometry*, 65(2): 331–364, 2021.

Seshadri, V. and West, B. J. Fractal dimensionality of lévy processes. *Proceedings of the National Academy of Sciences*, 79(14):4501–4505, 1982.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Simsekli, U., Sagun, L., and Gurbuzbalaban, M. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pp. 5827–5837. PMLR, 2019.

Simsekli, U., Sener, O., Deligiannidis, G., and Erdogdu, M. A. Hausdorff dimension, heavy tails, and generalization in neural networks. *Advances in Neural Information Processing Systems*, 33:5138–5151, 2020.

Solow, A. R. and Polasky, S. Measuring biological diversity. *Environmental and Ecological Statistics*, 1:95–103, 1994.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Willerton, S. Heuristic and computer calculations for the magnitude of metric spaces. *arXiv preprint arXiv:0910.5500*, 2009.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

A. Intrinsic dimensions and equalities

The following result has been proven by Meckes (2015), which related the magnitude dimension with the Minkowski dimension of compact subsets of Euclidean space.

Theorem A.1 (Corollary 7.4, (Meckes, 2015))**.** *If $X \subset \mathbb{R}^n$ is compact and if $\dim_{\mathrm{Mag}} X$ or $\dim_{\mathrm{Mink}} X$ exist, then both exist and $\dim_{\mathrm{Mag}} X = \dim_{\mathrm{Mink}} X$.*

Using the theorem above, one can use the magnitude dimension to approximate the Minkowski dimension.

Theorem A.2 (Heine-Borel)**.** *Let $A \subset \mathbb{R}^n$. Then A is compact if and only if A is both closed and bounded.*

Theorem A.3. *(Schweinhart, 2021) Let $X \subset \mathbb{R}^n$ be a bounded set. Then $\dim^0_{\mathrm{PH}} X = \dim_{\mathrm{Mink}} X$.*

B. Assumption H1

The following technical assumption has been introduced in previous work (Birdal et al., 2021; Simsekli et al., 2020) and it is necessary for the statement of the novel bound in Theorem 4.2 For any $\delta > 0$, define the fixed grid on $\mathbb{R}^d$ to be

$$G = \left\{ \left(\frac{(2j_i + 1)\delta}{2\sqrt{\delta}}, \ldots, \frac{(2j_d + 1)\delta}{2\sqrt{\delta}} \right) : j_i \in \mathbb{Z}, i = 1, \ldots, d \right\}. \tag{10}$$

The collection of centres of each ball is denoted by the set N_δ. Now, we define $N_\delta(S) = \{x \in N_\delta : B_d(x, \delta) \cap \mathcal{W}_{\mathcal{S}} \neq \emptyset\}$, where S is the training set, $B_d(x, \delta) \subset \mathbb{R}^d$ denotes the closed ball centered around $x \in \mathbb{R}^d$ of radius δ. Then $N_\delta(S)$, as defined, is the collection of centres of each ball that intersects $\mathcal{W}$.

H1: Let $\mathcal{Z}^\infty := (\mathcal{Z} \times \mathcal{Z} \times \mathcal{Z} \times \ldots)$ denote the countable product endowed with the product topology and let $\mathfrak{B}$ be the Borel σ-algebra generated by $\mathcal{Z}^\infty$. Let $\mathfrak{F}, \mathfrak{G}$ be the sub-σ-algebras of $\mathfrak{B}$, generated by the collections of random variables given by $\{L_S(w) : w \in \mathcal{W}, n \geq 1\}$ and $\left\{ \mathbb{1}\{w \in N_\delta : \delta \in \mathbb{Q}_{>0}, w \in G, n \geq 1\} \right\}$ respectively. There exists a constant $M \geq 1$, such that for any $A \in \mathfrak{F}$, $B \in \mathfrak{G}$, we have $\mathbb{P}[A \cap B] \leq M\mathbb{P}[A]\mathbb{P}[B]$.

This is known as the ψ-mixing condition (Bradley, 1983) and it is common in statistics for proving limit theorems. Smaller M indicates that the dependence of $L_S(w)$ on the training sample S is weaker.