神经网络中的度量空间大小与泛化 s

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

arXiv:2305.05611v1 [cs.LG] 9 May 2023

摘要

深度学习模型已在众多应用中取得显著成功，但其内部运行机制仍不为人所知。本研究旨在通过一种称为大小的新型拓扑不变量来量化深度神经网络的学习过程。大小是一种等距不变量；其性质是当前的研究热点，因为它能够编码度量空间的许多已知不变量。我们利用大小研究神经网络的内部表征，并提出一种确定其泛化能力的新方法。此外，我们从理论上建立了大小维数与泛化误差之间的联系，并通过实验表明，所提出的框架能够很好地指示后者。

1. 引言

深度神经网络（DNN）凭借出色的性能，已广泛应用于多种任务，包括计算机视觉 Xie et al. (2017)、自然语言处理 Vaswani et al. (2017) 和科学发现 J umper et al. (2021)。然而，最先进的 DNN 通常包含数百万至数十亿个参数，因此人类用户难以准确理解这些网络的内部运行机制。一个关键但尚未解决的问题是理解神经网络的泛化能力，即其在未见数据上表现良好的能力。近期文献提出了神经网络内在维数的多种概念 Birdal 等人（2021）；Dupuis 等人（2023）；Simsekli 等人（2020），并指出泛化最重要的因素是有效容量，而非参数数量。

在这项工作中，我们提出了一种衡量神经网络内在维数的新方法，该方法基于一种称为大小的新型拓扑不变量。与以往方法不同，大小概念简单易懂，而且计算复杂度可能更低。这一选择也在理论上得到了论证。大小是度量空间的一种等距不变量，其性质是数学研究的活跃领域，因为它能够编码几何测度论和积分几何中的许多重要不变量（Leinster，2013）。此外，大小的研究源于理论生态学，它能够刻画生态系统中的有效物种数。更广泛地说，研究已表明，大小因此能够编码空间中不同点的有效数量（Leinster，2013）。我们进一步基于这一思想，提出了一种利用有效模型数这一概念评估泛化误差的新方法，并将其作为有效容量。

尽管关于大小的理论研究已取得显著进展，但其相当多的特性仍有待发现，且若干问题尚未得到解答，尤其是关于其在机器学习中的实际应用。虽然近期研究（Adamer 等，2021年；Bunch 等，2021年）已开始建立大小量向量与机器学习应用之间的联系，但这一研究领域仍处于非常早期的阶段。

近期研究表明，在某些条件下，一个源自大小的概念——大小维数——与闵可夫斯基维数相同（Meckes（2015））。因此，我们可以从理论上证明，训练算法轨迹的泛化误差与上述度量空间的大小维数存在内在联系。这使我们得以对神经网络的学习过程形成新的认识。此外，我们的工作首次探索了大小维数这一概念；它具有坚实的理论基础，可用作神经网络内在维数的一种度量。

贡献。我们的工作首次将大小这一数学概念引入理论深度学习和神经网络泛化研究；我们认为，这是对新兴的拓扑机器学习领域的一项令人振奋且新颖的贡献（Hajij 等人，2022；Hensel 等人，2021）。在建立理论界限后，我们探究了多个实验设置下大小函数的变化，并比较了多个尺度下的大小值与测试准确率。此外，我们通过实验表明，大小与测试准确率之间存在联系。通过与持久同调维数建立新的联系，我们进而将所提出的度量与 Adams 等人（2020）提出的内在维数进行比较，并证明我们的方法在计算复杂度和可解释性方面更具优势。

[1]爱丁堡大学信息学院 [2]亥姆霍兹慕尼黑研究中心 [3]慕尼黑工业大学。通讯作者：Rayna Andreeva <r.andreeva@sms.ed.ac.uk>。

预印本。审稿中。

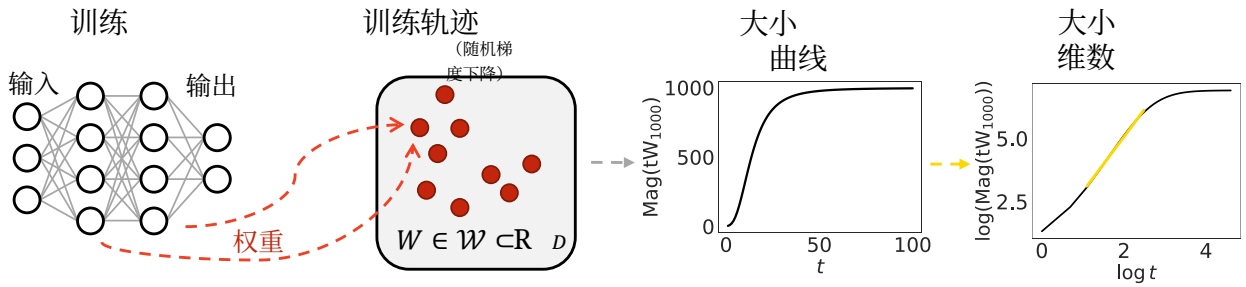


图*1*。 **流程概述**。 我们首先训练一个神经网络，并监测其训练轨迹。沿着轨迹每迭代 1000 次，我们就收集一次训练权重 $W \in \mathbb{R}^d$ 并将其组成一个点云。随后，我们计算该点云在选定尺度 t 下的幅度曲线，绘制双对数图，并据此估算大小维数。

由 Adams 等人（2020）提出，并证明我们的方法具有更好的计算复杂度和可解释性。简而言之，我们的贡献如下：

- 我们提出了一种基于大小和有效模型数来评估神经网络泛化能力的新方法，使我们无需验证集即可监测模型性能。
- 我们证明了一个新的泛化误差上界，将泛化误差与训练轨迹的一种基于大小的特征联系起来。
- 我们通过实验证明，这些度量在训练过程中的变化与测试集准确率相关。
- 我们证明，先前提出的所有内在维数概念都与大小维数相同，并通过实验证实了这一结果。

2. 相关工作

我们简要回顾与神经网络中的泛化、内在维数和大小相关的文献。

泛化界与内在维数。 一些研究探讨了内在维数，以刻画神经网络的泛化能力。Simsekli 等人（2020）证明，假设类的分形维数与泛化误差相关，而泛化误差又与网络轨迹的重尾行为有关（Hodgkinson & Mahoney, 2021; Mahoney & Martin, 2019; Simsekli 等人, 2019）。然而，要在实践中计算该界，需要满足许多假设。近期的一项研究放宽了部分假设，并提出了持久同调维数的概念（Adams 等人（2020）），$\dim_{\mathrm{PH}}$。Birdal 等人（2021）首次为使用拓扑不变量分析深度神经网络提供了理论依据。另一项研究（Magai & Ayzenberg, 2022）考察了网络不同深度和层上的 $\dim_{\mathrm{PH}}$，并观察了其演变过程。在 Dupuis 等人（2023）提出了一种数据驱动维度，并将其与 Birdal 等人（2021）的 $\dim_{\mathrm{PH}}$ 进行了比较。他们证明，该维度与泛化误差的相关性比以往研究所示的更强，并成功放宽了部分限制性假设。

大小及其在机器学习中的应用。Solow 与 Polasky（1994 年）首次提出大小，用于衡量生物多样性，但未涉及其数学性质。大约二十年后，Leinster（2013 年）才使用范畴论的语言，将其数学性质形式化。此后，Leinster 与 Shulman（2021 年）证明了大小同调中的大小可实现为欧拉示性数。尽管大小具有理论基础，但其在机器学习中的应用仍然很少。近年来，将大小引入机器学习领域重新受到关注。第一项在机器学习背景下发展大小概念的研究表明，大小的各个求和项（称为大小权重）可以用作高效的边界检测器（Bunch 等，2021 年）。此外，大小还被用于图像空间，并被证明是有效的边缘检测器（Adamer 等，2021 年）。然而，我们的工作首次将大小直接应用于深度学习。

利用拓扑刻画神经网络。早期研究已确立神经网络训练与拓扑不变量之间的联系（Fern´andez 等人，2021年），例如利用拓扑复杂度作为泛化性能的代理指标（Rieck 等人，2019年）。然而，这些研究仅分析完成训练过程后的已训练网络（Fern´andez 等人，2021年），可能忽略了训练动态的关键方面（Birdal 等人，2021年）。与此相反，我们提出使用另一种拓扑不变量——大小，它比以往方法具有更强的可解释性。此外，我们在训练轨迹上而非已训练网络上计算大小，从而提供关键的拓扑

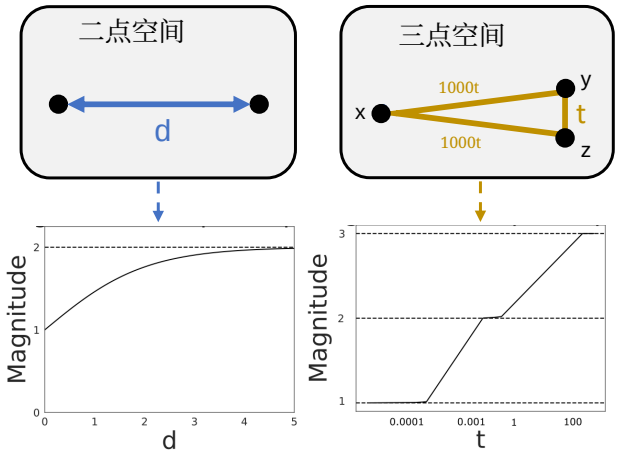


图 2。 **二点空间和三点空间的大小**。 左侧是二点空间，其中两点之间的距离为 d。右侧是等腰三角形的三点空间，其中 y 与 z之间的距离为 t， x 与 y 之间以及 x 与 z之间的距离为 $1000t$。在每个空间下方，我们可以看到相应的大小函数。

学见解，以帮助理解训练动态。

3. 背景

我们首先定义大小和大小维数，然后再定义内在维数。

3.1. 大小

虽然度量空间大小是一个一般性概念（Leinster，2013），但我们将重点限定在 $\mathbb{R}^n$ 的子集上，因为已知其大小存在（Meckes，2013）。

定义 3.1. 设 (X, d) 是一个有限度量空间，其距离度量为 d。那么， X 的相似度矩阵定义为 $\zeta_{ij} = e^{-d_{ij}}$，其中 $1 \leq i, j \leq n$，而 n 表示 X 的基数。

定义 3.2. 设 X 是一个相似度矩阵为 ζ_{ij} 的度量空间。如果 ζ_{ij} 可逆，则大小定义为

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

当 X 是 $\mathbb{R}^n$ 的有限子集时，正如 Leinster (2013)（定理 2.5.3）所证明的， ζ_{ij} 是一个对称正定矩阵。于是，$(\zeta^{-1})_{ij}$ 存在，因此大小也存在。考察几个点数较少的样本空间，最能说明大小的概念。

示例 3.3。 令 X 表示仅含一个点 a的度量空间。那么， ζ_X 是一个 1×1 矩阵，其中 $\zeta_X^{-1} = 1$，利用大小的公式，我们得到 $\mathrm{Mag}_X = 1$。

示例 3.4. 一个更直观的例子是由两个点构成的空间。设 $X = \{a, b\}$ 是一个有限度量空间，其中 $d_X(a, b) = d$。那么

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

使得

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

因此

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

图 2 也展示了这个例子。根据公式 (4)，如果 d 非常小，即 $d \to 0$，那么 $\mathrm{Mag}(X) \to 1$。类似地，当 $d \to \infty$ 时，$\mathrm{Mag}(X) \to 2$。实际上，从图 2 可以看出，即使 d 小至 5，$\mathrm{Mag}(X) = 2$。

通过考察度量空间的重缩放对应空间，可以获得关于该空间的更多信息。由此得到的表示更为丰富，并且可以用大小函数简洁地概括。为此，引入尺度参数 t。通过改变 t，我们可以得到度量空间的缩放版本，从而回答一个度量空间有多少个有效点。

大小函数。 大小为每个度量空间赋予的不只是一个标量，而是一个函数。具体做法如下：我们引入尺度参数 t，对于 t 的每个取值，考虑将点之间的距离按 t缩放后的空间。更具体地说，对于 $t > 0$，记号 tX 表示 X 按 t 倍缩放。计算 t 的每个取值对应的大小，就得到大小函数。更形式化地说，定义如下：

定义 3.5。 设 (X, d) 为一个有限度量空间。定义 (tX, d_t) 为一个度量空间，其点集与 X 相同，度量为 $d_t(x, y) = td(x, y)$。

这里的直观理解是，我们通过不同的尺度观察同一个空间，例如，将距离单位从米改为厘米。由此便自然得到大小函数的定义。

定义 3.6。 有限度量空间 (X, d) 的大小函数是部分定义函数 $t \to \mathrm{Mag}(tX)$，其对所有 $t \in (0, \infty)$ 都有定义。

示例 3.7。 考虑图 2 中三点空间的大小函数。在此例中，点 x、 y和 z 构成一个等腰三角形。当 $t = 0.0001$ 时，这三个点彼此非常接近，几乎无法区分。因此，我们说这个空间有 1 个有效点。相反，当 $t = 0.01$时，

右侧两点 y 和 z 之间的距离非常小，而 x 则相距很远。因此，我们说这个空间看起来像是由两点组成。最后，当 t 很大时，这三个点彼此相距很远，大小的值为 3，也就是这个空间的基数。

3.2. 内在维数

有多种概念可用于衡量空间的内在维数。在本文中，我们将重点介绍其中三种：上盒维数（闵可夫斯基维数）、大小维数和持久同调维数。盒维数以覆盖数为基础，并可通过 Simsekli 等人（2020）与泛化联系起来，而大小维数则建立在我们前面定义的概念之上。

定义 3.8（闵可夫斯基维数）。对于有界度量空间 X，令 $N_\delta(X)$ 表示中心位于 X 中且两两不相交的闭 δ-球的最大数量。上盒维数/闵可夫斯基维数定义为

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

这里有一点需要谨慎说明。一般而言，闵可夫斯基维数与豪斯多夫维数并不一致，也不等价。不过，在 Simsekli 等人（2020）中，作者给出了使我们所关注空间的豪斯多夫维数与闵可夫斯基维数一致的条件。事实上，对于许多类分形集合，这两种维数概念彼此相等；参见 Mattila（1999，第5章）。

定义 3.9（大小维数）。当

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

存在时，我们将其定义为X 的大小维数（Meckes, 2015）。

对于 t的某个适当取值区间，大小维数可近似理解为大小函数的变化率。我们还可以引入另一种分形维数概念，称为持久同调维数（$\dim_{\mathrm{PH}}$）（Adams 等人（2020））。

定义 3.10。有界度量空间（X，d）的持久同调维数记作 $\dim_{\mathrm{PH}}$，定义如下：

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

其中 E_α 是 α-寿命总和，定义为

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

b和 d 分别是 0 阶持久同调（PH^0）中的出生值和死亡值。它度量 W 的 Vietoris-Rips过滤中的所有连通分支，记为 $Rips(W)$。

这个定义相当技术化，对于本文的工作并非关键；更多详情请参阅 Adams 等人（2020）；Edelsbrunner 和 Harer（2022）；M´emoli 和 Singhal（2019）。

4. 理论结果

我们先阐明不同内在维数概念之间的联系，然后再证明它们与泛化误差之间的联系。

4.1. 内在维数概念之间的联系

介绍了三种不同的内在维数概念后，我们将证明，在一些温和的假设下，它们实际上是相同的。我们的新贡献是证明 $\dim_{\mathrm{Mag}} X$ 和 $\dim^0_{\mathrm{PH}} X$。该结果形式化为以下定理，并假设所有维数概念均存在。

定理 4.1. 设 $X \subset \mathbb{R}^n$ 为紧集，且维数$_{\mathrm{Mag}} X$ 或维数$_{\mathrm{Mink}} X$ 存在。那么

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

证明。由于 X 是紧集，且 $\dim_{\mathrm{Mag}} X$ 或 $\dim_{\mathrm{Mink}} X$ 存在，根据推论 7.4 和 Meckes（2015），可知 $\dim_{\mathrm{Mag}} X$ 和 $\dim_{\mathrm{Mink}} X$ 均存在，且 $\dim_{\mathrm{Mag}} X = \dim_{\mathrm{Mink}} X$。由于 X 是紧集，根据海涅–博雷尔定理，X 既是闭集又是有界集，因此根据 Kozma 等（2006）；Schweinhart（2021）中的一个结果，有 $\dim_{\mathrm{Mink}} X = \dim^0_{\mathrm{PH}} X$。因此，$\dim_{\mathrm{Mag}} X = \dim_{\mathrm{Mink}} X = \dim^0_{\mathrm{PH}} X$，这意味着 $\dim_{\mathrm{Mag}} X = \dim^0_{\mathrm{PH}} X$。

□

4.2. 与泛化误差的联系

在确立了各种维数概念之间的等价性之后，我们将着手形式化机器学习理论所需的语言，并最终得出一项新颖的泛化结果。

本节开头，我们沿用 Shalev-Shwartz 与 Ben-David（2014）的符号。为使本文内容自成一体，我们简要回顾一些标准定义。在标准的统计学习设定中，$\mathcal{X}$ 表示特征集合，$\mathcal{Y}$ 表示标签集合。二者的笛卡尔积

乘积 $\mathcal{X} \times \mathcal{Y}$ 表示数据空间 $\mathcal{Z}$。学习者可以访问由 m 个样本组成的数据序列，称为训练数据，记作 $S = ((x_1, y_1), ..., (x_m, y_m)) \text{ in} \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$。我们假设训练集 S 由 $\mathcal{X}$ 上的某个未知概率分布生成，将其记作 $\mathcal{D}$，且每个样本 $\{x_i, y_i\}$ 都是来自$\mathcal{D}$的独立同分布（i.i.d.）样本。我们将专注于一个受限的搜索空间，以寻找一组预测器，并将其称为假设类 $\mathcal{H}$。其中的每个元素称为假设，记作 $h \in \mathcal{H}$，是一个从 $\mathcal{X}$到 $\mathcal{Y}$的函数。通过计算一个称为损失函数的量来选出最优假设或参数向量 $h \in \mathcal{H}$，该损失函数定义为 $\ell \colon \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$。于是，对于 $z \in \mathcal{Z}$，经验误差为$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z)$，而真实误差或风险为 $L_{\mathcal{D}}(h) = \mathbb{E}_z[\ell(h, z)]$，其中 $z \in \mathcal{Z}$。泛化误差定义为真实误差与经验风险之差，即 $|L_S(h) - L_{\mathcal{D}}(h)|$。

对于神经网络，假设类为$\mathcal{W} \subset \mathbb{R}^d$，每个 $w \in \mathcal{W}$ 都是一个参数向量。给定训练算法 $\mathcal{A}$，我们希望研究针对给定训练数据 S，优化过程 $\mathcal{A}$ 所返回的所有假设类的集合。我们将其称为优化轨迹，并记为 $\mathcal{W}$。为获取该过程每一步的迭代结果，我们用 w_i 表示第 i次迭代时 $\mathcal{W}$ 中的一个元素。更具体地说，$\mathcal{W} := \{w \in \mathbb{R}^d : \exists i \in [0, I], \ w = [\mathcal{A}(S)_i]\}$，其中 I 是训练迭代次数。换句话说，当固定i时，我们关注优化算法 $\mathcal{A}$在第 i次迭代返回的权重。

对于下面的结果，Birdal 等人（2021，假设 H1）提出了一个要求更为严格的技术条件，详见附录。我们将该假设记为 H1。我们要求存在常数 $M > 0$，用于量化集合 $\mathcal{W}_S$ 对训练样本 S的依赖程度。现在我们已经给出了所有必要的定义，可以继续介绍这一新结果。

定理 4.2. 设 $\mathcal{W} \in \mathbb{R}^d$ 为一个紧集。假设 $H1$ 成立，则 ℓ 由常数 C界定，并且关于 w是 K-*Lipschitz* 连续的。对于充分大的 n，我们有如下界：

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

概率为 $1 - \gamma$ *over* $S \sim \mathcal{D}^{\otimes n}$，其中 M 是假设 *H1* 中的常数。

证明。由于 $\mathcal{W}$ 有界，我们有 $\dim_{\mathrm{PH}}^0\mathcal{W}$ =$\dim_{\mathrm{Mag}}\mathcal{W}$。因此，将命题1（Birdal 等人，2021年）中的 $\dim_{\mathrm{PH}}^0\mathcal{W}$ 替换为 $\dim_{\mathrm{Mag}}\mathcal{W}$ 即可得出结果。 □

5. 方法

与以往关于内在维数的研究不同，我们提出使用大小来估计分形维数。正如我们所见，大小维数的概念与闵可夫斯基维数和持久同调维数一致。这一理论联系使我们能够有信心地在神经网络背景下探索大小的概念。具体做法如下：在权重轨迹 $\mathcal{W}$上的选定点处，我们对若干模型 W进行子采样，这些模型可视为一个点云，其中每个点都是模型空间中的一个模型。该空间的维数非常高，等于网络中的参数数量。我们计算每个此类点云的大小和大小维数。随后，我们研究这些量与测试准确率之间的联系。

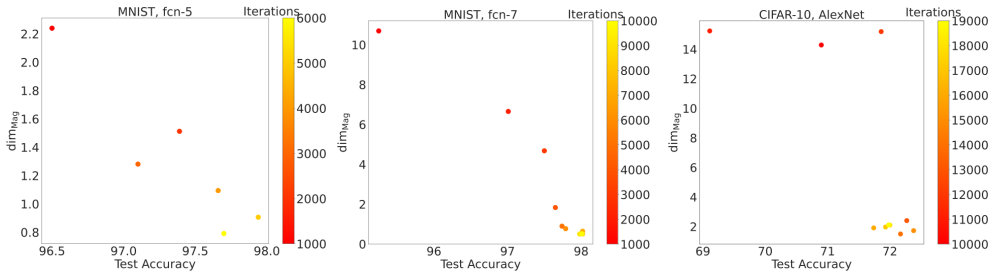
鉴于我们在定理 4.2 中得出的新结果将内在维数与泛化界联系起来，自然会引出一个问题：大小函数是否也可用于探索神经网络的学习过程？由于大小维数可以粗略地解释为大小函数的变化率，因此，如果大小维数与泛化误差之间存在联系，那么不同尺度下大小函数的取值也应与泛化误差相关。我们想要研究的是，随着学习过程推进，模型轨迹空间如何变化。换言之，随着训练的进行，这个空间看起来是包含更多彼此不同的点，还是更像是包含较少的点？用大小的语言来表达这个想法，在进行更多次学习算法的迭代时，有效模型数 是在增加还是减少？

由于大小是一个函数，我们希望对幅度曲线进行横截面切片，并考察在特定尺度参数 t 选择下的大小值。因此，我们给出了固定 t时有效模型数的定义。

定义5.1。我们将**有效模型数**定义为$\mathrm{Mag}(t_i\mathcal{W})$ 在尺度 t_i 处的值。

5.1. 通过幅值维数分析深度神经网络动力学

估计大小维数并非一项简单的任务，因为需要通过找出大小函数中最长的直线部分来近似公式6中的极限；这一过程无法自动计算，必须手动完成。这里我们给出从有限样本 W 中计算 $\dim_{\mathrm{Mag}}\mathcal{W}$ 的算法，以近似无限过程。我们采用了与 O'Malley 的大小维数估计类似的步骤



(a) (b) (c)

图*3*。**大小维数与测试准确率相关**。在图(a–c)中，我们绘出了大小维数与测试准确率的关系，并展示了不同迭代次数下的结果，颜色由红到黄表示。我们注意到，在 MNIST 和 CIFAR-10 数据集以及不同架构（FCN-5、FCN-7、AlexNet）中，大小维数与测试准确率均存在相关性，且这种相关性在 MNIST 上强于 CIFAR-10。

算法1 $\dim_{\mathrm{Mag}}\mathcal{W}$ 的估计

输入： 训练轨迹集合 $\mathcal{W} = \{w_i\}^n$ 大小为 n，尺度参数 t： $[0, t_k]$，区间 $[t_i, t_j]$，$i < j < k$

输出： $\dim^{\mathrm{Mag}}$

对于 $r = 1$**到** k **执行**

$\mathrm{Mag}(t_r\mathcal{W}) \leftarrow \mathcal{W}$

结束循环

$m, b \leftarrow$ 拟合直线($log(\mathrm{Mag}(\mathcal{W}_n)[t_i : t_j])$ $log([t_i : t_j])$)

维度$_{\mathrm{Mag}}\mathcal{W} \leftarrow m$

等人（2023）；Willerton（2009）。具体步骤见算法1。选定合适的代表区间 $[t_i, t_j]$ 后，绘制大小与 t 的双对数图，并计算所选区间 $[t_i, t_j]$ 上直线的斜率 m 和截距 b。然后，将 m 作为 $\dim_{\mathrm{Mag}}\mathcal{W}$ 的估计值。从本质上说，这一过程是基于斜率计算极限，也就是应用洛必达法则。

6. 实验结果

本节的首要目标是验证我们的核心主张，即我们对大小维数的估计能够衡量泛化能力。第二个目标是建立大小本身与测试准确率之间的联系。第三个目标是对两个非常接近的分形维数度量，即 $\dim_{\mathrm{Mag}}\mathcal{W}$ 和 $\dim_{\mathrm{PH}}\mathcal{W}$，进行实证比较。第四个目标是尝试解释我们的研究结果对于神经网络中的泛化问题究竟意味着什么，并由此得出新的见解。为了展示这一点，我们将所提出的方法应用于多种不同的神经网络架构、训练设置和数据集。具体而言，我们使用不同范围的学习率和100的批量大小，分别在 MNIST 数据集上训练5层网络（FCN-5）和7层网络（FCN-7），并在 CIFAR10 数据集上训练 AlexNet（Krizhevsky 等，2017）。

我们考虑一个包含1000次训练迭代的滑动窗口。然后，我们按照算法1中的步骤估计每个窗口的 $\dim_{\mathrm{Mag}}\mathcal{W}$。

6.1. 探索学习过程

我们评估了本文的主要论断，即大小维数与泛化误差之间存在联系。图3揭示了多项发现。首先，我们观察到大小维数与测试准确率之间存在相关性——大小维数越低，测试准确率越高。这一结果与 Birdal 等人（2021）和 Simsekli 等人（2020）此前观察到的现象一致，尽管其视角来自于大小这一不变量；与持久同调相比，大小在实践中通常更易于计算，也更具可解释性。其次，这一结论同时适用于 MNIST 和 CIFAR10，也适用于不同的架构。这表明，尽管参数存在显著差异，不同数据集的内在维数仍可能相近。为了实现良好的泛化性能，重要的是在不因将重要的表征特征压缩到同一维数而丢失这些特征的前提下，使维数尽可能小。

6.2. 网络轨迹的分析与可视化

接下来，我们想研究随着网络经过更多迭代次数的训练，大小函数会受到怎样的影响。为此，我们从$[0, 40]$至 $t \in \{1.36, 6.78, 16.95, 30.51\}$ 的范围内选取四个 t 值，并计算模型的有效数量 $\mathrm{Mag}(t\mathcal{W})$，以初步了解训练轨迹的空间。我们进一步固定学习率，并改变迭代次数。为便于说明，实验采用 0.1 的学习率。不过，我们注意到，这种模式在多种学习率下都成立。图 4 展示了所选 t尺度下得到的大小值，颜色表示迭代次数。我们注意到

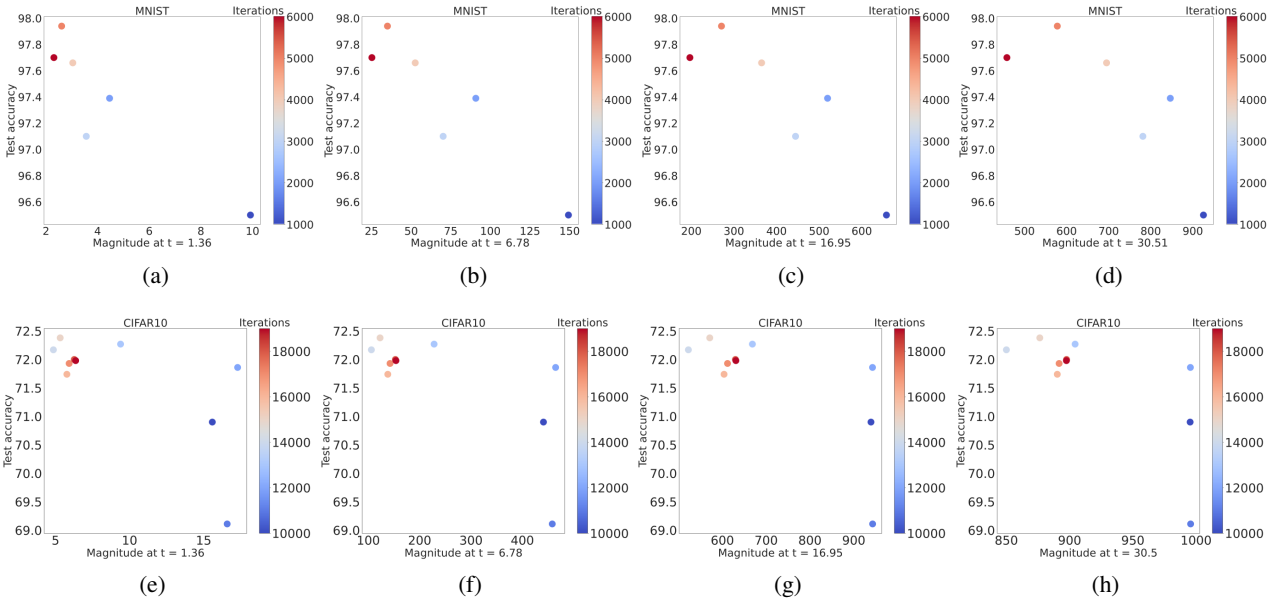


图 *4.* **有效模型数与测试准确率相关**。 这里展示了在不同 t 值下对幅度曲线进行的横截面评估。每个点表示模型轨迹经过 1000 次迭代的情况，幅度是在此过程中计算的，同时还展示了滑动窗口中最后一个模型的测试准确率。第一行（图 (a-d)）展示了 MNIST 的大小。第二行（图 (e-h)）展示了 CIFAR10 的结果。我们注意到，在所有尺度下，测试准确率与大小值之间的相关性模式相似。

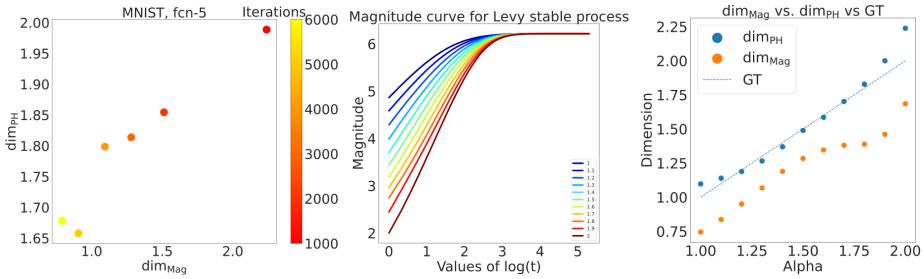


图 *5.* α**-Levy 稳定过程的大小维数、持久同调维数和真实值**。 图 (a) 比较了 $\dim_{\mathrm{Mag}}\mathcal{W}$ 和 $\dim_{\mathrm{PH}}\mathcal{W}$ 与迭代次数之间的关系。二者具有很强的相关性，相关系数为 0.96，且 p 值具有统计显著性（$p < 0.05$）。图 (b) 展示了不同 α 值下的大小曲线。图 (c) 展示了多个 α 值下 α- 稳定 Levy 过程的大小曲线。

左上角聚集着红色点，表明较高的测试准确率对应较低的大小值。同样，对于蓝色点，较高的大小值与较差的测试集性能相关。这种模式在 MNIST 和 CIFAR10 上均成立。此外，测试准确率越低，大小值越高，这意味着大小值较低的模型泛化能力更好。这一结果直观上是合理的。当大小值较小时，这个空间看起来包含的点更少。随着 t增加，大小会收敛到该集合的基数。在（a）中，测试准确率较高时，有效模型数约为 2

率，这是一个有趣的现象。这意味着，在 1000 个模型中，对于 $t = 1.38$，这个空间看起来像是由 2 个模型构成的，这表明训练轨迹形成了 2 个大的簇。令人惊讶的是，对于经过 1000 次迭代训练、测试准确率最低的网络，情况并非如此。在这种情况下，模型空间看起来像是由 10 个点构成的，而且更加分散。

请注意，即使在更大的尺度下，测试准确率较高的模型大小也更小，这表明模型呈现出某种聚类行为。用大小的术语来说，不同模型的有效数量

在网络泛化能力更强时会更少；“良好的优化轨迹”的大小低于“糟糕的优化轨迹”；学习算法通过将模型聚类为少数几个簇，隐式地优化了大小。在整个实验中，我们观察到大小本身与测试准确率之间存在很强的相关性，而且这种相关性在不同尺度、学习率和数据集下均持续存在。因此，图 4 结果的主要见解是，较少的有效模型数量有利于泛化，并表明存在一种聚类模式。

6.3. $\dim_{\mathrm{Mag}}\mathcal{W}$ 与$\dim_{\mathrm{PH}}\mathcal{W}$之间的相似性

在图 5(a) 中，我们展示了 MNIST 上 $\dim_{\mathrm{Mag}}\mathcal{W}$ 与 $\dim_{\mathrm{PH}}\mathcal{W}$ 之间存在很强的相关性。尽管二者之间有很好的对应关系，但 $\dim_{\mathrm{PH}}\mathcal{W}$ 更难解释。为了对其作出一些解释，作者不得不退后一步，生成用于计算指定维数的持久图。然而，由于他们必须对空间进行采样以估计维数，这些计算涉及大量不同的图，因此，将 $\dim_{\mathrm{PH}}\mathcal{W}$ 与原始轨迹联系起来并不那么直接。另一方面，大小、大小维数与权重参数簇的出现之间存在更清晰的联系。

消融研究。在图 5 中，我们展示了消融研究的结果。我们使用一个 d维的 α-稳定 Lévy 过程 Seshadri & West (1982) 来模拟一个与权重轨迹相似、真实值已知的过程，其中取 $d = 10$，并将大小维数的值与分形维数进行比较。此外，我们还将其与持久同调维数进行比较；正如我们在定理 4.1 中证明的那样，持久同调维数与大小维数相同。本研究旨在通过实证评估这一理论结果。如图 5(b) 所示，我们的维数与真实值高度相关。它似乎始终比真实维数低约 0.3。尽管如此，这种高度显著的统计相关性表明，对于一个类似于随机梯度下降迭代过程的过程，大小维数确实非常接近分形维数。这让我们非常确信，大小维数测量的正是它应当测量的内容。

7. 结论

我们的工作首次建立了大小与泛化误差之间的联系。我们证明了一个理论结果，将优化轨迹的大小维数与测试准确率联系起来。我们通过探索多种实验设置和数据集中的大小变化，在实验中验证了我们的结果。我们进一步加深了对权重轨迹聚类特性的理解：通过理论和实证结果，我们证明，与测试表现较差、大小更大且分布更分散的模型相比，泛化误差较低的模型往往聚集得更紧密。Birdal 等人（2021）；Br¨uel-Gabrielsson 等人（2019）此前描述过这一现象，而我们的观察为其提供了补充。未来，我们将改进大小维数的估计，并研究将大小用作正则化器。此外，我们证明，仅凭大小本身，无需验证集便可监控神经网络的表现。未来，我们可以研究将大小用作早停准则，类似于 Rieck 等人（2019）。此外，从我们的实证消融研究来看，大小维数似乎始终低于真实值，这一点还需进一步研究。具体而言，我们希望探究在多大程度上可以改进定理 4.2 中的泛化界。

致谢

作者谨向 AlexandrosKeros 和 Emily Roff 致谢，感谢他们提出有益的意见并进行富有见地的讨论，帮助改进了本文。RA 获英国研究与创新（资助编号 EP/S02431X/1）、爱丁堡大学 UKRI 生物医学人工智能博士培训中心、信息学院以及国际亥姆霍兹—爱丁堡表观遗传学研究学院（EpiCrossBorders）的支持。KL 获亥姆霍兹协会通过联合研究生院“慕尼黑数据科学学院 - MUDS”的支持。

参考文献

Adamer, M. F., DeBrouwer, E., O’Bray, L. 和 Rieck, B. 图像的幅度向量。*arXiv* 预印本 *arXiv:2110.15188*，2021年。

Adams, H., Aminian, M., Farnell, E., Kirby, M., Mirth, J., Neville, R., Peterson, C. 和 Shonkwiler, C. 一种基于持久同调的测度分形维数。载于拓扑数据分析：*2018* 年阿贝尔研讨会，第 1–31 页。施普林格，2020年。

Birdal, T., Lou, A., Guibas, L. J., and Simsekli, U. 神经网络中的内在维数、持久同调与泛化。神经信息处理系统进展，34，2021年。

Bradley, R. C. 关于平稳随机序列的 ψ-混合条件。美国数学会汇刊，276(1):55–66，1983年。

Br¨uel-Gabrielsson, R., Nelson, B. J., Dwaraknath, A., Skraba, P., Guibas, L. J., and Carlsson, G. 一种用于机器学习的拓扑层。*arXiv* 预印本 *arXiv:1905.12200*，2019年。

Bunch, E.、Kline, J.、Dickinson, D.、Bhat, S. 和 Fung, G. 面向机器学习的加权向量：应用于边界检测的数值调和分析。*arXiv* 预印本 *arXiv:2106.00827*，2021年。

Dupuis, B.、Deligiannidis, G. 和 S¸ims¸ekli, U. 具有数据依赖分形维数的泛化界。*arXiv* 预印本 *arXiv:2302.02766*，2023年。

Edelsbrunner, H. 和 Harer, J. L. 计算拓扑：入门。美国数学学会，2022年。

Fern´andez, D. P.、Guti´errez-Fandi˜no, A.、Armengol-Estap´e, J. 和 Villegas, M. 使用代数拓扑确定人工神经网络的结构性质。*arXiv* 预印本 *arXiv:2101.07752*，2021年。

Hajij, M., Zamzmi, G., Papamarkou, T., Miolane, N., Guzm´an-S´aenz, A., Natesan Ramamurthy, K., Birdal, T., Dey, T. K., Mukherjee, S., Samaga, S. N., Livesay, N., Walters, R., Rosen, P. 和 Schaub, M. T. 拓扑深度学习：超越图数据。*arXiv* 预印本，文章编号 arXiv:2206.00606，2022年。doi：10.48550/arXiv.2206. 00606。

Hensel, F., Moor, M. 和 Rieck, B. 拓扑机器学习方法综述。人工智能前沿，4:681108，2021年。

Hodgkinson, L. 和 Mahoney, M. 随机优化中的乘性噪声与重尾。载于国际机器学习会议，第 4262–4274 页。PMLR，2021年。

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., ˇZ´ıdek, A., Potapenko, A. 等。利用 AlphaFold 进行高精度蛋白质结构预测。*Nature*，596(7873):583–589，2021年。

Kozma, G., Lotker, Z. 和 Stupp, G. 最小生成树与上盒维数。*Proceedings ofthe AmericanMathematical Society*，134(4):1183–1187，2006年。

Krizhevsky, A., Sutskever, I. 和 Hinton, G. E. 使用深度卷积神经网络进行 ImageNet 分类。*Communications ofthe ACM*，60(6):84–90，2017年。

Leinster, T. 度量空间的大小。*DocumentaMathematica*，18:857–905，2013。

Leinster, T. and Shulman, M. Magnitude homology of enrichedcategoriesandmetric 空间。*Algebraic & Geo-metric Topology*，21(5):2175–2221，2021年。

Magai, G. 和 Ayzenberg, A. 深度学习中数据流形的拓扑与几何。*arXiv preprintarXiv:2204.08624*，2022年。

Mahoney, M. 和 Martin, C. 神经网络模型中的传统自正则化与重尾自正则化。载于国际机器学习会议，第 4284–4293 页。PMLR，2019。

Mattila, P. 欧几里得空间中集合与测度的几何：分形与可求长性。第 44 卷。剑桥大学出版社，1999 年。

Meckes, M. W. 正定度量空间。《*Positivity*》，17(3)：733–757，2013。

Meckes, M. W. 度量空间的大小、多样性、容量与维数。《*Potential Analysis*》，42(2)：549–572，2015。

M´emoli, F. 和 Singhal, K. 有限度量空间持续同调入门。《数学生物学通报》，81：2074–2116，2019。

O'Malley, M.、Kalisnik, S. 和 Otter, N. α-大小。《纯粹与应用代数杂志》，第 107396 页，2023。

Rieck, B., Togninalli, M., Bock, C., Moor, M., Horn, M., Gumbsch, T., 和 Borgwardt, K. 神经持久性：使用代数拓扑衡量深度神经网络复杂度的复杂度量。载于国际学习表征会议 *(ICLR)*，2019年。网址：https: //openreview.net/forum?id=ByxkijC5FQ。

Schweinhart, B. 持久同调与上盒维数。离散与计算几何，65(2)：331–364，2021年。

Seshadri, V. 和 West, B. J. Lévy 过程的分形维数。美国国家科学院院刊，79(14)：4501–4505，1982年。

Shalev-Shwartz, S. 和 Ben-David, S. 理解机器学习：从理论到算法。剑桥大学出版社，2014年。

Simsekli, U.、Sagun, L. 和 Gurbuzbalaban, M. 深度神经网络中随机梯度噪声的尾指数分析。载于国际机器学习会议，第 5827–5837 页。PMLR，2019 年。

Simsekli, U., Sener, O., Deligiannidis, G. 和 Erdogdu, M. A. 豪斯多夫维数、重尾与神经网络中的泛化。神经信息处理系统进展，33:5138–5151，2020年。

Solow, A. R. 和 Polasky, S. 测量生物多样性。环境与生态统计，1:95–103，1994。

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., 和 Polosukhin, I. 注意力机制就是你所需要的一切。神经信息处理系统进展，30，2017。Willerton, S. 度量空间大小的启发式方法和计算机计算。*arXiv* 预印本 *arXiv:0910.5500*，2009。Xie, S., Girshick, R., Doll´ar, P., Tu, Z., 和 He, K. 深度神经网络的聚合残差变换。载于*IEEE*计算机视觉与模式识别会议论文集，第1492–1500页，2017。

A. 内在维度与等式

以下结果由 Meckes（2015）证明，该结果将大小维数与欧几里得空间紧子集的闵可夫斯基维数联系起来。

定理 A.1（推论 7.4，Meckes（2015））。 如果 $X \subset \mathbb{R}^n$ 是紧的，且维数$_{\mathrm{Mag}}X$ 或维数$_{\mathrm{Mink}}X$ 存在，那么二者都存在，并且维数$_{\mathrm{Mag}}X =$ 维数$_{\mathrm{Mink}}X$。

利用上述定理，可以使用大小维数来近似闵可夫斯基维数。

定理 A.2（海涅–博雷尔定理）。 设 $A \subset \mathbb{R}^n$。当且仅当 A 既闭又有界时， A 是紧的 *ed.*

定理 A.3。 *Schweinhart*（*2021*）设 $X \subset \mathbb{R}^n$ 为有界集合。那么维数$^0_{\mathrm{PH}}X =$ 维数$_{\mathrm{Mink}}X$。

B. 假设 H1

以下技术假设由先前的工作 Birdal 等人（2021）； Simsekli 等人（2020）提出，是陈述定理 4.2 中新界限所必需的。对于任意 $\delta > 0$，定义 $\mathbb{R}^d$ 上的固定网格为

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

每个球的中心所构成的集合记为 N_δ。现在，我们定义 $N_\delta(S) = \{x \in N_\delta :\ B_d(x,\ \delta) \cap \mathcal{W}_S \neq \emptyset\}$，其中 S 是训练集， $B_d(x,\ \delta) \subset \mathbb{R}^d$ 表示以 $x \in \mathbb{R}^d$ 为中心、半径为 δ的闭球。于是，按定义， $N_\delta(S)$是与 $\mathcal{W}$相交的每个球的中心所构成的集合。

H1：令 $\mathcal{Z}^\infty \coloneqq (\mathcal{Z} \times \mathcal{Z} \times \mathcal{Z} \times \ldots)$表示赋有积拓扑的可数乘积，并令 $\mathfrak{B}$ 为由 $\mathcal{Z}^\infty$生成的 Borel σ-代数。令 $\mathfrak{F}$， $\mathfrak{G}$ 分别为 $\mathfrak{B}$的子-σ-代数，由以下随机变量族生成： $\{L_S(w) :\ w \in \mathcal{W},\ n \geq 1\}$ 以及 $\left\{ \mathbb{1}\{w \in N_\delta :\ \delta \in \mathbb{Q}_{>0},\ w \in G,\ n \geq 1\} \right\}$。存在一个常数$M \geq 1$，使得对于任意 $A \in \mathfrak{F}$、 $B \in \mathfrak{G}$，均有 $\mathbb{P}[A \cap B] \leq M\mathbb{P}[A]\mathbb{P}[B]$。

这被称为 ψ-混合条件（Bradley，1983年），在统计学中常用于证明极限定理。较小的M 表示 $L_S(w)$对训练样本 S 的依赖性较弱。