Zero Generalization Error Theorem for Random Interpolators via Algebraic Geometry

Anonymous authors
Paper under double-blind review

Abstract

We theoretically demonstrate that the generalization error of random interpolators in teacher-student regression problems achieves zero if the number of training samples exceeds a certain threshold. Understanding why highly over-parameterized models generalize well remains a central problem in machine learning theory. While recent studies have often attributed this phenomenon to the implicit bias of optimization algorithms such as stochastic gradient descent (SGD), empirical evidence indicates that this phenomenon primarily stems from properties of the model itself. Specifically, even randomly sampled interpolators, namely parameters that achieve zero training error, have been observed to generalize effectively. In this paper, we study the generalization behavior of random interpolators in noiseless teacher-student regression problems. We prove that the generalization error of a randomly sampled interpolator becomes exactly zero once the number of training samples exceeds a threshold determined by the geometric structure of the interpolator set in parameter space. Technically, we leverage tools from algebraic geometry to characterize this geometric structure.

1 Introduction

A central question in modern learning theory is why large models such as deep neural networks (DNNs) achieve excellent generalization performance despite perfectly fitting the training data (Simonyan & Zisserman, 2015; Zhang et al., 2017). Such models achieving zero training error are commonly referred to as *interpolators*. Explaining the performance of these interpolators constitutes one of the central challenges in contemporary deep learning theory, and several distinct lines of work have emerged to address this phenomenon (Belkin et al., 2018; Muthukumar et al., 2020; Bartlett et al., 2020; Liang & Rakhlin, 2020; Liang et al., 2020; Belkin, 2021).

Among the lines of work for interpolators, one of the leading explanations has been the *implicit bias* induced by learning algorithms, typically stochastic gradient descent (SGD) and its variants. This perspective posits that SGD guides parameters toward generalizing solutions in parameter space among many possible ones. For example, in stylized settings, gradient-based methods are known to converge to solutions with certain norms, margins, or structural properties (Gunasekar et al., 2017; Soudry et al., 2018; Arora et al., 2019; Lyu & Li, 2020; Chizat & Bach, 2020; Yun et al., 2021). Despite substantial progress, this line of explanation faces two important limitations. First, both theoretical and empirical studies have reported cases where implicit bias does not necessarily enhance generalization (Dauber et al., 2020; Vyas et al., 2024; Farhang et al., 2022). Second, most of the existing theoretical analyses are restricted to simplified settings, such as linear models or two-layer neural networks, due to a fundamental difficulty in identifying the solution for non-convex optimization problems.

In parallel, a growing body of recent work has suggested that the generalization behavior of interpolators may be explained by *model-based properties* that are determined by the architecture and geometry of the model itself, rather than the implicit bias of optimization. In particular, several empirical studies have observed that predictors obtained by SGD can resemble predictors sampled from suitable distributions over interpolating solutions (Valle-Pérez et al., 2019; Mingard et al., 2021). Related empirical evidence indicates

that SGD-trained parameters and randomly sampled interpolating parameters in DNNs can exhibit similar generalization performance (Chiang et al., 2023).

An important challenge within this line of research is to provide a theoretical foundation for the empirical observations above. Specifically, we pose the following question:

Can the strong generalization performance of interpolators be explained by a model-based theory that does not rely on the implicit bias of algorithms?

1.1 Our Result

In this study, we develop a model-based theory that explains the generalization performance of random interpolators of large models. Specifically, in noiseless teacher-student regression, we demonstrate that random interpolators attain exactly zero generalization error with a finite number of training samples. We call the minimum such number the *strong sample complexity*, and we derive an explicit upper bound for it.

Informally, our main theorem states the following.

Theorem 1 (Informal statement of Theorem 2)**.** *The following holds with probability* 1*:*

(Strong sample complexity) $\leq$ *(Dimension of parameter space)* $-$ *(Dimension of teacher-equivalent set)* $+ 1$.

Here, the *teacher-equivalent set* is the set of parameters of the student model that replicate the teacher model for every input in the input space; see Section 3.2 for a formal definition. This result shows that the strong sample complexity can be substantially smaller than the total number of parameters when the teacher-equivalent set has large dimension, as is the case for DNNs; see Theorems 4 and 5. It therefore provides a partial theoretical explanation for why DNNs can generalize well in the over-parameterized regime, where the total number of parameters far exceeds the number of training samples.

Technically, our analysis relies on the concept of *real analytic sets*, originating in algebraic geometry, defined as intersections of zero sets of analytic functions in real space. Specifically, we utilize real analytic sets for studying the relationship between the geometric structure of the interpolator set and that of the teacher-equivalent set, since interpolators are precisely characterized as the zeros of the loss function evaluated on the training data.

We summarize our contributions as follows:

1. We propose a model-based theory of generalization for random interpolators, focusing on the geometry of the interpolating set rather than on the implicit bias of a particular optimization algorithm.
2. In noiseless teacher-student regression, we derive an explicit upper bound on the strong sample complexity required for random interpolators to achieve zero generalization error; see Theorem 2.
3. We apply the theory to fully connected DNNs and ResNets and show that the resulting strong sample complexity bound can be substantially smaller than the number of parameters; see Theorems 4 and 5. This result partially explains why over-parameterized DNNs generalize well, which classical theory fails to explain.
4. We introduce the concept of real analytic sets from algebraic geometry into learning theory for analyzing the relationship between interpolation, parameter redundancy, and exact generalization.

1.2 Related Work

We review previous theoretical studies on the generalization ability of interpolators, organized by the model class considered in each line of work. We then discuss related geometric analyses of interpolator sets and parameter redundancy.

Linear regression and kernel regression. A substantial body of work has studied interpolation in linear and kernel models. Belkin et al. (2018) showed that interpolation can be compatible with good

generalization, thereby challenging the classical view that exact fitting necessarily leads to overfitting. For over-parameterized linear regression, Muthukumar et al. (2020) studied exact interpolation of noisy data, while Bartlett et al. (2020) characterized when the minimum-norm interpolator exhibits benign overfitting. For kernel methods, Liang & Rakhlin (2020) studied ridgeless kernel regression, and Liang et al. (2020) analyzed the risk of minimum-norm interpolants in reproducing kernel Hilbert spaces. Relatedly, Belkin (2021) argued that over-parameterized kernel interpolators can generalize due to an implicit bias toward simple functions.

Random feature models and two-layer neural networks. Several works have analyzed interpolation beyond linear models via random feature models and two-layer neural networks. Yang et al. (2021) studied uniform generalization bounds for interpolating predictors in random feature models and showed a gap between classical bounds and actual generalization behavior. Mei & Montanari (2022) analyzed random feature regression, which can be viewed as a two-layer network with random first-layer weights, and derived precise asymptotics exhibiting double descent.

Deep neural networks. Theoretical analyses of interpolators in deep neural networks with more than two layers remain relatively limited both in classification and regression settings. The work most closely related to ours is Buzaglo et al. (2024), which developed a model-based theory for random interpolators of quantized deep neural networks in binary classification. They showed that typical random interpolators can generalize when the labels are generated by a narrow teacher network. Our work differs in that we study noiseless teacher-student regression and prove exact zero generalization error after finitely many samples. Moreover, our strong sample complexity bound is expressed in terms of the dimension of the parameter space and the dimension of the teacher-equivalent set. Yu et al. (2024) also studied interpolation in deep networks through memorization and derived sample complexity for classification problems.

Geometric structure of interpolator sets. Our work is also related to studies of the geometry of the set of interpolating parameters. Cooper (2021) studied the loss landscape of over-parameterized neural networks and related the number of training samples to the dimension of the set of global minima. Fukumizu et al. (2019) analyzed how a narrower teacher network can be embedded into a wider student network while preserving the same input-output function, which is closely related to our construction of teacher-equivalent sets in Theorems 4 and 5.

Identifiability. The teacher-equivalent set studied in this paper is closely related to the identifiability of neural networks. Classical and recent works have studied conditions under which two neural networks represent the same function only through finitely many parameter choices, typically up to natural symmetries such as permutations of hidden units (Sussmann, 1992; Fefferman, 1994; Vlačić & Bölcskei, 2022; Usevich et al., 2025; Shahverdi et al., 2026). Our setting is complementary: we focus on over-parameterized student models in which the teacher-equivalent set can have positive dimension, and we show that this redundancy can reduce the strong sample complexity. Moreover, whereas these prior works have suggested a close connection between parameter redundancy and the generalization behavior of deep neural networks, our results provide a theoretical perspective supporting this connection.

Other related work. Our use of real analytic geometry is related in spirit to singular learning theory (Watanabe, 2009), which applies algebraic-geometric tools to singular or non-identifiable statistical models. While singular learning theory mainly studies the asymptotic behavior of Bayesian inference, our work uses real analytic sets to prove finite-sample exact generalization of random interpolators in over-parameterized models.

2 Preliminaries

2.1 Notation

Let $\mathbb{N} = \{1, 2, \ldots\}$. For $n \in \mathbb{N}$, we write $[n] := \{1, 2, \ldots, n\}$. We denote the sets of non-negative real numbers and non-negative integers by $\mathbb{R}_{\geq 0}$ and $\mathbb{Z}_{\geq 0}$, respectively. For a (finite-dimensional) vector, we write $\|\cdot\|$ for

the Euclidean norm and $\|\cdot\|_1$ for the 1-norm. For a nonempty set $S \subset \mathbb{R}^d$ and a point $\omega \in \mathbb{R}^d$, we denote the distance from ω to S by $\|\omega - S\| := \inf_{s \in S} \|\omega - s\|$. For a third-order tensor $w \in \mathbb{R}^{m \times M \times M}$, we denote by $w_{i,:,:} \in \mathbb{R}^{M \times M}$ the slice of w obtained by fixing the first index to i and allowing the remaining indices to vary. For a second-order tensor $h \in \mathbb{R}^{m \times M}$, we denote $h_{i,:}$ in the same way.

2.2 Noiseless Teacher-Student Regression

We formalize the noiseless teacher-student regression setting studied in this paper.

Data-generating process. Let $\mathcal{X} \subset \mathbb{R}^m$ be the input space and let $\mathcal{Y} \subset \mathbb{R}^p$ be the output space. We assume that $\mathcal{X}$ is a compact and connected real analytic manifold of dimension m, possibly with boundary; the definition of a real analytic manifold is given in Section 3.2.1. We also assume that $\mathcal{Y}$ is a p-dimensional set. The output corresponding to an input $x \in \mathcal{X}$ is generated by a fixed teacher model $f^*(\cdot; \theta^*) : \mathcal{X} \to \mathcal{Y}$:

$$y = f^*(x; \theta^*),$$

where $\theta^* \in \mathbb{R}^{d^*}$ is a fixed teacher parameter with its dimension $d^* \in \mathbb{N}$.

We observe n training samples $\{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$, where $x_1, \ldots, x_n$ are drawn independently according to a probability distribution $\mathcal{D}$ on $\mathcal{X}$, and

$$y_i = f^*(x_i; \theta^*) \quad \text{for each} \quad i \in [n].$$

This data-generating setting with a noiseless teacher is commonly used in theoretical analyses of neural networks (Tian, 2017; Safran & Shamir, 2018; Xu & Du, 2023).

For technical reasons, we impose the following assumption on the distribution $\mathcal{D}$ of the input data x.

Assumption 1 (Data distribution)**.** *The probability distribution $\mathcal{D}$ of an input x is absolutely continuous with respect to the uniform distribution on $\mathcal{X}$.*

Student model and losses. We define a student model which is learned by using the samples. Given the training samples, we consider a student model

$$f(x; \theta), \ x \in \mathcal{X}, \theta \in \Theta,$$

where $\Theta \subset \mathbb{R}^{d_\Theta}$ is the student parameter space and θ is an $\mathbb{R}^{d_\Theta}$-valued parameter to be learned. In addition, $\Theta \subset \mathbb{R}^{d_\Theta}$ is a compact d_Θ-dimensional real analytic manifold, possibly with boundary.

We use the squared loss $\ell(y, y') := \frac{1}{2}\|y - y'\|^2$. The training loss and the generalization error are defined by

$$L_n(\theta) := \frac{1}{n} \sum_{i=1}^n \ell\big(y_i, f(x_i; \theta)\big), \qquad L(\theta) := \mathbb{E}_{x \sim \mathcal{D}} \left[\ell\big(f^*(x; \theta^*), f(x; \theta)\big)\right].$$

2.3 Real Analytic Function

We introduce an important concept for our analysis, a real analytic function.

Definition 1 (Real analytic function)**.** Let U be an open subset of $\mathbb{R}^d$ for $d \in \mathbb{N}$. A function $f : U \to \mathbb{R}$ is called real analytic if for every point $\theta^{(0)} \in U$, f can be expressed as a convergent power series which converges in a neighborhood of $\theta^{(0)}$:

$$f(\theta) = \sum_{j=0}^{\infty} \sum_{\alpha : \|\alpha\|_1 = j} c_\alpha (\theta - \theta^{(0)})^\alpha,$$

where $c_\alpha \in \mathbb{R}$ is a constant, $\theta = (\theta_1, \cdots, \theta_d) \in U$ and $\alpha = (\alpha_1, \cdots, \alpha_d) \in \mathbb{Z}_{\geq 0}^d$ is a multi-index of non-negative integers where we define $\theta^\alpha = \theta_1^{\alpha_1} \cdots \theta_d^{\alpha_d}$. When the output of f is multi-dimensional, f is called a real analytic function if the mapping from the input to each of the output components is a real analytic function.

Throughout our analysis, we put the following real analytic assumption.

Assumption 2 (Real analytic student and teacher model)**.** *We assume that $f(x;\theta)$ is a real analytic function with respect to $\theta \in \Theta$ and $x \in \mathcal{X}$. Furthermore, we assume that $f^*(x;\theta^*)$ is a real analytic function with respect to $x \in \mathcal{X}$.*

A wide range of machine learning models satisfy this assumption: for example, a fully connected deep neural network or an attention mechanism whose activation function is analytic, such as sigmoid, softmax, and hyperbolic tangent.

3 Interpolator and Teacher-Equivalent Set

We introduce key concepts that form the foundation of our analysis. The first is a predictor based on a *random interpolator*, with particular emphasis on implementations using neural networks. The second is a *teacher-equivalent set*, a notion that plays a central role in our generalization theory.

3.1 Predictor with Random Interpolator

We consider a predictor that interpolates the training data by sampling a parameter from a distribution supported on the set of parameters that perfectly interpolate the training data.

In preparation, we consider an *interpolating parameter set* (IPS), satisfying zero training error, *i.e.*,

$$\widehat{\Theta}_n := \{\theta \in \Theta \mid L_n(\theta) = 0\} = \{\theta \in \Theta \mid \ell(y_i, f(x_i;\theta)) = 0, \forall i \in [n]\}. \tag{1}$$

Note that $\widehat{\Theta}_n$ is a random set from the random training samples $\{(x_i, y_i)\}_{i=1}^n$.

Next, we define $\mathbb{P}(\cdot \mid \widehat{\Theta}_n)$ as a probability distribution that is absolutely continuous with respect to the uniform distribution on $\widehat{\Theta}_n$ (the mathematically precise definition is given in Appendix A.1). We then consider a predictor with the sampled parameter:

$$f(x;\widehat{\theta}_n), \quad \widehat{\theta}_n \sim \mathbb{P}(\cdot \mid \widehat{\Theta}_n). \tag{2}$$

This random predictor interpolates the training samples with probability 1, i.e., $\mathbb{P}(L_n(\widehat{\theta}_n) = 0) = 1$. We note that, since Θ is compact, the random interpolator is sampled exclusively from this compact set, even when the set of all interpolators is not compact.

3.2 Teacher-Equivalent Set (TES)

We define the notion of a teacher-equivalent set (TES). We say that a parameter $\theta \in \Theta$ of a student model is *teacher-equivalent* when $f(x;\theta) = f^*(x;\theta^*)$ holds for every $x \in \mathcal{X}$. Then, the *teacher-equivalent set* (TES) is the set of teacher-equivalent parameters:

$$\bar{\Theta} := \{\theta \in \Theta \mid f(x;\theta) = f^*(x;\theta^*), \forall x \in \mathcal{X}\}.$$

Intuitively, the TES is the set of student parameters that realize exactly the same function as the teacher model. Unlike the IPS $\widehat{\Theta}_n$, the TES $\bar{\Theta}$ is deterministic and does not depend on the training samples.

In our analysis, we assume that the TES is not empty as a realizability condition.

Assumption 3 (Realizability)**.** *The TES $\bar{\Theta}$ is non-empty, i.e., there exists a parameter $\theta^\circ \in \Theta$ satisfying $f^*(x;\theta^*) = f(x;\theta^\circ)$ for every $x \in \mathcal{X}$.*

This realizability assumption is natural for the teacher-student setup with smaller teachers (Tian, 2017; Safran & Shamir, 2018; Xu & Du, 2023). In fact, this assumption is satisfied in the case of fully connected DNNs and ResNets, as we see later in Section 5.

As an important observation, the TES $\bar{\Theta}$ can be considered to be the population version of the IPS $\widehat{\Theta}_n$. Therefore, as the training sample size n diverges to infinity, a student model from $\widehat{\Theta}_n$ interpolates all possible data points in the sample space $\mathcal{X} \times \mathcal{Y}$, and thus $\widehat{\Theta}_n$ asymptotically converges to $\bar{\Theta}$. Indeed, as n increases, $\widehat{\Theta}_n$ monotonically shrinks and approaches $\bar{\Theta}$, as we see in Section 4. Figure 1 illustrates this intuition.

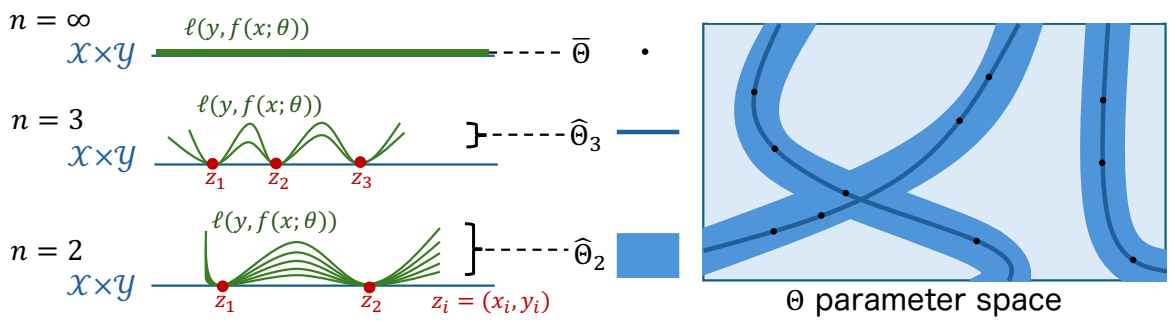

Figure 1: Illustration of the IPS $\widehat{\Theta}_n$ approaches the TES $\bar{\Theta}$. As n increases, the dimension of $\widehat{\Theta}_n$ decreases, and thus $\widehat{\Theta}_n$ converges to $\bar{\Theta}$. In the figure, we illustrate $\widehat{\Theta}_2$ as 2-dimensional, $\widehat{\Theta}_3$ as 1-dimensional, and $\bar{\Theta}$ as 0-dimensional (i.e., a finite set of points).

3.2.1 Dimension of TES

We define the dimension of the TES $\bar{\Theta}$. In preparation, we define a real analytic manifold and its dimension.

Definition 2 (Real analytic manifold)**.** Let $\Omega \subset \mathbb{R}^d$. We say that Ω is a d'-dimensional real analytic manifold if there exists an atlas $\{(U_i, \varphi_i)\}_i$ such that $\{U_i \cap \Omega\}_i$ covers Ω, each U_i is open in $\mathbb{R}^d$, and $\varphi_i : U_i \cap \Omega \to V_i$ is a homeomorphism onto an open subset $V_i \subset \mathbb{R}^{d'}$. Moreover, whenever $U_i \cap U_j \cap \Omega \neq \varnothing$, the transition map $\varphi_j \circ \varphi_i^{-1} : \varphi_i(U_i \cap U_j \cap \Omega) \to \varphi_j(U_i \cap U_j \cap \Omega)$ is real analytic. In the case of a manifold with boundary, we replace the condition $V_i \subset \mathbb{R}^{d'}$ by $V_i \subset \mathbb{H}^{d'} := \{x \in \mathbb{R}^{d'} : x_{d'} \geq 0\}$, and require the transition maps to extend real analytically to open neighborhoods in $\mathbb{R}^{d'}$.

We remark that in the Euclidean space $\mathbb{R}^d$, all the open subsets are real analytic manifolds in the sense of standard topology and their dimensions correspond to d.

We now state the definition of the dimension of the TES.

Definition 3 (Dimension of TES)**.** The dimension of the TES $\bar{\Theta}$, denoted as $d_{\bar{\Theta}}$, is defined as the maximum dimension of a real analytic manifold Ω contained in $\bar{\Theta}$.

Intuitively, the dimension of the TES measures the degrees of freedom among student parameters that realize the teacher model. For example, if there exists a unique parameter $\theta' \in \Theta$ that is teacher-equivalent, then the TES is a singleton and hence $d_{\bar{\Theta}} = 0$. Such a situation occurs in linear regression models, as we see later in Section 4.1. In contrast, singular models including deep neural networks typically have a positive-dimensional TES due to parameter redundancy, as we discuss later in Section 5.

4 Generalization Error Analysis

We analyze the generalization error of a predictor with a random interpolator defined in (2).

4.1 Main Theorem

We investigate the number of training samples required for a predictor with a random interpolator to achieve zero generalization error. To facilitate the analysis, we define the important notion, the *strong sample complexity*, which quantifies the required number of training samples.

Definition 4 (Strong sample complexity for a random interpolator)**.** For a predictor with a random interpolator defined in (2), its strong sample complexity is defined by

$$\widehat{k} := \inf \left\{ n \in \mathbb{N} \mid L(\widehat{\theta}_n) = 0 \text{ a.s. for } \widehat{\theta}_n \sim \mathbb{P}(\cdot \mid \widehat{\Theta}_n) \right\},$$

where $\widehat{\Theta}_n$ is a random set in terms of $\mathcal{D}$ defined in (1).

The strong sample complexity is a random variable in terms of $\mathcal{D}$ which represents the minimum necessary number of training samples for $\widehat{\theta}_n$ to achieve zero generalization error almost surely. This notion is a stronger version of the ordinary sample complexity, which is the required number of training samples for achieving a generalization error smaller than some positive value.

We now show our main result on the generalization error of a predictor with a random interpolator. We provide its proof in Appendix C.

Theorem 2 (Strong sample complexity: general case)**.** *Suppose that Assumptions 1, 2, and 3 hold. Then, the strong sample complexity for a random interpolator satisfies the following with probability one over training samples:*

$$\widehat{k} \leq d_{\Theta} - d_{\bar{\Theta}} + 1.$$

Theorem 2 shows that the strong sample complexity is determined by the difference between the dimension of the parameter space and the dimension of the TES. In particular, a larger TES dimension reduces the number of training samples required for a random interpolator to achieve zero generalization error almost surely.

We next illustrate this result through several examples. First, consider linear regression with a linear teacher $f^*(x;\theta^*) = {\theta^*}^\top x$ and a linear student $f(x;\theta) = \theta^\top x$, where $\theta^*, \theta \in \mathbb{R}^m$. In this case, the TES is the singleton $\bar{\Theta} = \{\theta^*\}$ and hence, $d_{\bar{\Theta}} = 0$. Therefore, Theorem 2 yields

$$\widehat{k} \leq m + 1.$$

This is consistent with the classical observation that noiseless linear regression requires a number of samples comparable to the parameter dimension in order to recover the teacher model.

In contrast, fully connected deep neural networks (FCDNNs) typically have highly redundant parameterizations, which lead to TESs of large dimension, as shown later in Section 5. Consequently, the strong sample complexity can be substantially smaller than the parameter dimension:

$$\widehat{k} \ll d_{\Theta}.$$

This observation is qualitatively consistent with the empirical phenomenon that deep neural networks often generalize well even in highly over-parameterized regimes.

Remark 1 (Comparison with standard generalization analyses)**.** Classical uniform-convergence-based analyses typically yield nonzero upper bounds on the generalization error that decrease with the number of training samples, often at a rate of order $O(1/\sqrt{n})$. In contrast, our analysis shows that, in the noiseless teacher-student setting, random interpolators achieve exactly zero generalization error once the number of training samples exceeds the strong sample complexity. This difference arises because our analysis focuses on parameters that achieve strictly zero training error, which is not considered in the standard setting.

4.2 Extension to Near Interpolator Case

To address more practical scenarios, we extend our analysis to parameters that are not exact interpolators but are close to the interpolator set.

For $\varepsilon > 0$, define the ε-neighborhood of the interpolator set $\widehat{\Theta}_n$ by $\widehat{\Theta}_{n,\varepsilon} := \{\theta \in \Theta \mid \|\theta - \widehat{\Theta}_n\| \leq \varepsilon\}$. Then, we define a predictor with a *random near interpolator* as follows:

$$f(x;\widehat{\theta}_{n,\varepsilon}), \quad \widehat{\theta}_{n,\varepsilon} \sim \mathbb{P}(\cdot \mid \widehat{\Theta}_{n,\varepsilon}),$$

where $\mathbb{P}(\cdot \mid \widehat{\Theta}_{n,\varepsilon})$ denotes the uniform distribution with respect to the Lebesgue measure on $\widehat{\Theta}_{n,\varepsilon}$.

The following proposition shows that near interpolators also achieve small generalization error. Its proof is deferred to Appendix D.

Proposition 3 (Generalization error of near interpolators)**.** *Fix $\varepsilon > 0$ sufficiently small. Suppose that Assumptions 1, 2, and 3 hold and the number of training samples satisfies $n \geq d_\Theta - d_{\bar{\Theta}} + 1$. Moreover, we assume that there exists a universal constant $q > 0$ such that for every $x \in \mathcal{X}$, the map $\theta \mapsto f(x;\theta)$ is q-Lipschitz continuous. Then, with probability one over the training samples $Z_n = (x_1, \cdots, x_n)$, there exist constants $C(Z_n) < \infty$ and $\varepsilon_0(Z_n) > 0$ such that for all $0 < \varepsilon < \varepsilon_0(Z_n)$,*

$$L(\widehat{\theta}_{n,\varepsilon}) \leq (q\varepsilon)^2$$

holds with probability at least $1 - C(Z_n)\varepsilon$ in terms of $\widehat{\theta}_{n,\varepsilon} \sim \mathbb{P}(\cdot \mid \widehat{\Theta}_{n,\varepsilon})$.

The Lipschitz continuity assumption is standard in machine learning theory and is satisfied by many practical architectures, including fully connected neural networks, convolutional neural networks, and transformers with analytic activation functions. We note that Proposition 3 is not intended to characterize the low-training-loss parameters, but to provide a sufficient guarantee for the particular near-interpolator distribution $\mathbb{P}(\cdot \mid \widehat{\Theta}_{n,\varepsilon})$.

4.3 Proof Outline of Theorem 2 with Real Analytic Sets

We prove Theorem 2 by showing that $\widehat{\Theta}_n \setminus \bar{\Theta}$ becomes a null set with respect to the (intrinsic) Lebesgue measure on $\widehat{\Theta}_n$ when $n \geq d_\Theta - d_{\bar{\Theta}} + 1$. This is established by showing that the geometric dimension of $\widehat{\Theta}_n \setminus \bar{\Theta}$ becomes less than that of $\bar{\Theta}$, using tools from the theory of real analytic sets. Once this fact is established, it follows that $\ell(y, f(x;\widehat{\theta}_n)) = 0$ for every $x \in \mathcal{X}$ and for almost every $\widehat{\theta}_n \in \widehat{\Theta}_n$, which directly implies $\mathbb{E}[\ell(y, f(x;\widehat{\theta}_n))] = 0$.

In preparation, we introduce the key notion, a real analytic set. The IPS $\widehat{\Theta}_n$ is a real analytic set from the definition.

Definition 5 (Real analytic set)**.** A set of zeros of real analytic functions is called a *real analytic set*; that is, for real analytic functions $f_1, \cdots, f_n : \Theta \to \mathbb{R}$, a real analytic set is defined as

$$\{\theta \in \Theta \mid f_1(\theta) = 0, \cdots, f_n(\theta) = 0\}.$$

We further define the dimension of a real analytic set, which is the most important concept in our analysis.

Definition 6 (Dimension of a real analytic set)**.** The dimension of a real analytic set A is defined as the maximum d' such that A contains a real analytic manifold of dimension d'.

Intuitively, the dimension of a real analytic set is the number of degrees of freedom among variables that satisfies the equations. We note that from the definition, the dimension of a real analytic set is an integer. As an important remark, we define the dimension of $\widehat{\Theta}_n \setminus \bar{\Theta}$ by the maximum dimension of a real analytic manifold contained in $\widehat{\Theta}_n \setminus \bar{\Theta}$, analogously to 6.

An important property of real analytic sets is that they can be treated locally in much the same way as algebraic varieties, i.e., sets defined as the common zeros of polynomials. This stems from the fact that an analytic function can be locally expressed as a convergent power series, that is, as a sum of polynomials. Moreover, analogous results from complex algebraic geometry have been established via the technique of complexification, developed by Whitney & Bruhat (1959). As a consequence, many results concerning algebraic varieties also hold for real analytic sets.

Among the most fundamental results concerning the dimensions of algebraic varieties is Krull's principal ideal theorem (Hartshorne, 1977). Informally, Krull's principal ideal theorem states that the dimension of the solution set of n polynomial equations decreases by at least n under suitable conditions. For example, let $\theta = (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3$ and consider the solution set of two equations $\theta_1 + \theta_2 = 0$ and $\theta_2\theta_3 = 0$. The resulting set is $\{(a, -a, 0) \mid a \in \mathbb{R}\} \cup \{(0, 0, a) \mid a \in \mathbb{R}\}$, which consists of two straight lines in $\mathbb{R}^3$, and hence has dimension 1.

By applying arguments analogous to Krull's principal ideal theorem to real analytic sets, the dimension of $\widehat{\Theta}_n \setminus \bar{\Theta}$ becomes less than or equal to $d_\Theta - n$, since $\widehat{\Theta}_n$ is defined as a solution set of n equations. Moreover,

because $\widehat{\Theta}_n \supset \bar{\Theta}$ and the dimension of $\bar{\Theta}$ is $d_{\bar{\Theta}}$, it follows that the dimension of $\widehat{\Theta}_n \setminus \bar{\Theta}$ is strictly smaller than that of $\bar{\Theta}$ whenever $n \geq d_\Theta - d_{\bar{\Theta}} + 1$. Consequently, $\widehat{\Theta}_n \setminus \bar{\Theta}$ becomes a null set with respect to the Lebesgue measure on $\widehat{\Theta}_n$, which establishes the theorem.

Remark 2 (Mathematical novelty)**.** Our main technical contribution is to show that the dimension of the zero set decreases by the number of equations, *with probability one* in terms of the distribution $\mathcal{D}$ of the samples $x_1, \ldots, x_n$. Previously, it was known that such a dimension reduction event occurs under appropriate regularity conditions, such as the constant-rank condition (Theorem 11.2, Tu (2008)) or the transversality condition (Lemma 4.1.2, Robbin & Salamon (2025)). However, since verifying such regularity conditions is generally difficult, it is unclear whether they hold in our case. To overcome this issue, we prove that the dimension reduction event occurs with probability one with respect to $\mathcal{D}$, without requiring verification of such regularity conditions. This argument corresponds to Step 1 and Step 2 in the proof of the main theorem in Section C.

5 Application to Deep Neural Networks

We apply Theorem 2 to deep neural networks by analyzing the dimension of the teacher-equivalent set (TES). In particular, we derive upper bounds on the strong sample complexity through lower bounds on the TES dimension.

5.1 Fully Connected Deep Neural Networks

We study fully connected deep neural networks (FCDNNs) with analytic activation functions.

Teacher-Student Setup. The student model is an L-layer FCDNN with parameters $w^{(\ell)} \in \mathcal{W}^{(\ell)}$ and $b^{(\ell)} \in \mathcal{B}^{(\ell)}$, where $\mathcal{W}^{(\ell)} \subset \mathbb{R}^{m_\ell \times m_{\ell-1}}$ and $\mathcal{B}^{(\ell)} \subset \mathbb{R}^{m_\ell}$ are sufficiently large compact sets for $\ell = 1, ..., L$. The network is defined by

$$f(x;\theta) = w^{(L)}\sigma(w^{(L-1)} \cdots \sigma(w^{(1)}x + b^{(1)}) \cdots + b^{(L-1)}) + b^{(L)}.$$

We write $\theta = \{w^{(\ell)}, b^{(\ell)}\}_{\ell=1}^{L}$. Here, σ denotes a real analytic activation function. Moreover, $m_0 = m$ and $m_L = p$ denote the input and output dimensions, respectively.

The teacher model is also assumed to be an L^*-layer FCDNN with the same activation function σ. Its parameters are denoted by $\theta^* = \{{w^{(\ell)}}^*, {b^{(\ell)}}^*\}_{\ell=1}^{L^*}$, where ${w^{(\ell)}}^* \in \mathbb{R}^{m_\ell^* \times m_{\ell-1}^*}$ and ${b^{(\ell)}}^* \in \mathbb{R}^{m_\ell^*}$.

We assume that the student network is at least as wide as the teacher network:

$$L = L^*,\ m_\ell \geq m_\ell^*\ (1 \leq \ell \leq L^* - 1).$$

Result. We now present our result on the strong sample complexity of random interpolators for FCDNNs. The proof is provided in Appendix E.

Theorem 4 (Strong sample complexity for FCDNNs)**.** *Suppose that Assumption 1 holds, together with the teacher-student setting described above. Then the strong sample complexity satisfies*

$$\widehat{k} \leq m_1^*(m+1) + \sum_{\ell=2}^{L-1}(m_\ell^*(m_{\ell-1}+1)) + p(m_{L-1}+1) + 1$$

almost surely with respect to the training samples. In terms of the student model size, this scales as

$$\widehat{k} = O\left(\sum_{\ell=1}^{L-1} m_\ell\right).$$

Theorem 4 shows that the strong sample complexity grows linearly with the network width, whereas the total number of parameters grows quadratically with width. Thus, even in highly over-parameterized regimes, the number of training samples required for exact generalization can remain substantially smaller than the total parameter count.

5.2 Residual Networks

Teacher-Student Setup. We next study residual networks (ResNets) (He et al., 2016), which are convolutional neural networks with skip connections. To avoid unnecessarily complicating the theoretical analysis while still enabling discussion of the essential aspects, we analyze a simplified one-dimensional ResNet analyzed by Oono & Suzuki (2019).

The student model consists of N residual blocks, each of which has depth-L and width-M convolutional layers with size-K kernels. It is defined by a mapping $f(x;\theta)$ such that

$$f = \mathrm{FC}_{w,b} \circ \mathrm{Flatten} \circ \mathrm{Block}_{W^{(N)},B^{(N)}} \circ \cdots \circ \mathrm{Block}_{W^{(1)},B^{(1)}} \circ \mathrm{Padding},$$

where each factor is defined in the following:

- $\mathrm{Padding} : \mathbb{R}^m \to \mathbb{R}^{m\times M}$ denotes a zero padding operation defined by $\mathrm{Padding}(x) := (\underbrace{x\ 0\ \dots\ 0}_{M})$ for $x \in \mathbb{R}^m$.
- $\mathrm{Block}_{W^{(i)},B^{(i)}} : \mathbb{R}^{m\times M} \to \mathbb{R}^{m\times M}$ denotes a residual block consisting of convolutional layers of depth L with parameters $W^{(i)} = \{w^{(i,\ell)}\}_{\ell=1}^{L}$, $B^{(i)} = \{b^{(i,\ell)}\}_{\ell=1}^{L}$, where $w^{(i,\ell)} \in \mathbb{R}^{K\times M\times M}$ and $b^{(i,\ell)} \in \mathbb{R}^M$:
$$\mathrm{Block}_{W^{(i)},B^{(i)}} := \mathrm{Conv}_{w^{(i,L)},b^{(i,L)}} \circ \cdots \circ \mathrm{Conv}_{w^{(i,1)},b^{(i,1)}} + \mathrm{id},$$
where id denotes an identity mapping.
- $\mathrm{Conv}_{w^{(i,\ell)},b^{(i,\ell)}} : \mathbb{R}^{m\times M} \to \mathbb{R}^{m\times M}$ is a convolution with a size-K filter and an activation function σ,
$$\mathrm{Conv}_{w^{(i,\ell)},b^{(i,\ell)}}(h)_{k,:} := \sigma\left(\sum_{p=1}^{K} w^{(i,\ell)}_{p,:,:} h_{k+p-1,:} + b^{(i,\ell)}\right),$$
for $h \in \mathbb{R}^{m\times M}$ and $k = 1, \cdots, m$, with a one-sided padding defined by $h_{i,:} = 0$ for $i = m+1, \cdots, m+K-1$.
- $\mathrm{Flatten} : \mathbb{R}^{m\times M} \to \mathbb{R}^{mM}$ denotes the concatenation of column vectors defined by $\mathrm{Flatten}(h) := (h_{:,1}^\top \ \cdots \ h_{:,M}^\top)^\top$ for $h \in \mathbb{R}^{m\times M}$.
- $\mathrm{FC}_{w,b} : \mathbb{R}^{mM} \to \mathbb{R}^p$ denotes a fully connected layer with an identity activation with parameters $w \in \mathbb{R}^{p\times mM}$ and $b \in \mathbb{R}^p$, defined by $\mathrm{FC}_{w,b}(h) := wh + b$ for $h \in \mathbb{R}^{mM}$.

The teacher model is the same form of a ResNet with N^* residual blocks, each of which has depth-L^* and width-M^* convolutional layers with size-K kernels.

We assume that the student model is over-parameterized relative to the teacher model in terms of the number of residual blocks and width, that is,

$$N \geq N^*,\ M \geq M^*, L = L^*.$$

Moreover, the activation function σ is shared by the student and teacher networks and is assumed to be real analytic and satisfy $\sigma(0) = 0$.

Result. We now present the strong sample complexity bound for random interpolators in ResNets. The proof is provided in Appendix F.

Theorem 5 (Strong sample complexity for ResNets)**.** *Suppose that Assumption 1 holds, together with the teacher-student setting described above. Then the strong sample complexity satisfies*

$$\widehat{k} \leq N^* L M^*(KM+1) + (N-N^*)M^*(KM+1) + (mM+1)p + 1$$

almost surely with respect to the training samples. In terms of the student model size, this scales as

$$\widehat{k} = O(LKM + NKM).$$

Theorem 5 shows that the strong sample complexity grows linearly with the width M, the depth L, and the number of residual blocks N. In contrast, the total number of trainable parameters grows quadratically with the width, on the order of $NLKM^2$. Thus, even highly over-parameterized ResNets can achieve exact generalization with fewer training samples than the total number of parameters.

5.3 Proof Outline

By Theorem 2, it suffices to construct a positive-dimensional subset of the teacher-equivalent set. This is achieved by exploiting parameter redundancy in over-parameterized architectures.

For FCDNNs, the additional hidden units introduced by over-parameterization can be made functionally inactive by setting appropriate parameters to zero. As a result, the student network can exactly reproduce the teacher network while retaining redundant degrees of freedom in parameter space.

For ResNets, the additional residual blocks can similarly be made inactive by setting the parameters in the final layer of each additional residual block to zero. In this case, the identity skip connections propagate the teacher representation through the network, allowing the student model to reproduce the teacher function exactly while preserving positive-dimensional redundancy.

6 Experiment

We present a synthetic teacher-student regression experiment using fully connected deep neural networks (FCDNNs) designed to validate the theoretical predictions.

Specifically, we investigate how the generalization error of the near interpolators introduced in Section 4.2 depends on the number of training samples under the same teacher-student setting as in Section 5.1, where both the teacher and student models are FCDNNs.

Data and model setup. We consider a regression task with the mean squared error (MSE) loss. The input $x \in \mathbb{R}^2$ is sampled uniformly from $[-1, 1]^2$.

The student models are FCDNNs with depths 2, 4, and 6, each having 10 hidden units per hidden layer. For each student model, the corresponding teacher model has the same depth and consists of 5 hidden units per hidden layer. Both the teacher and student networks use the hyperbolic tangent activation function.

The output y is given by the scalar output of the teacher network, whose parameters are randomly initialized using Xavier uniform initialization (Glorot & Bengio, 2010).

Sampling setup. To sample near interpolators, we employ the *Guess and Check* (G&C) algorithm, which provides a practical implementation to sample random near interpolators (Chiang et al., 2023) (see Appendix G.1 for details). Specifically, we repeatedly sample parameters at random according to Xavier uniform distribution until the training loss falls below 0.01, and collect 1000 near interpolators for each training sample size.

We evaluate the test loss using 2000 independently generated test samples from the corresponding teacher network.

We employ G&C instead of stochastic gradient descent (SGD) because our theory concerns distributions over interpolators that are not tied to a specific optimization algorithm. In contrast, SGD may introduce optimization-dependent biases toward particular regions of the parameter space.

Results. Figure 2 shows the relationship between the test loss of random near interpolators and the number of training samples for FCDNNs of different depths. According to Theorem 4, the theoretical upper bounds on the strong sample complexity for the three student models are 27, 137, and 247, respectively. These values are indicated by the red vertical lines in Figure 2.

Across all three models, the transition in test loss occurs near the theoretically predicted sample scale, consistent with the upper bound in Theorem 4. This result suggests that the bound captures the onset of generalization for near interpolators.

As predicted by Proposition 3, the test loss does not converge exactly to zero because the sampled models are near interpolators rather than exact interpolators.

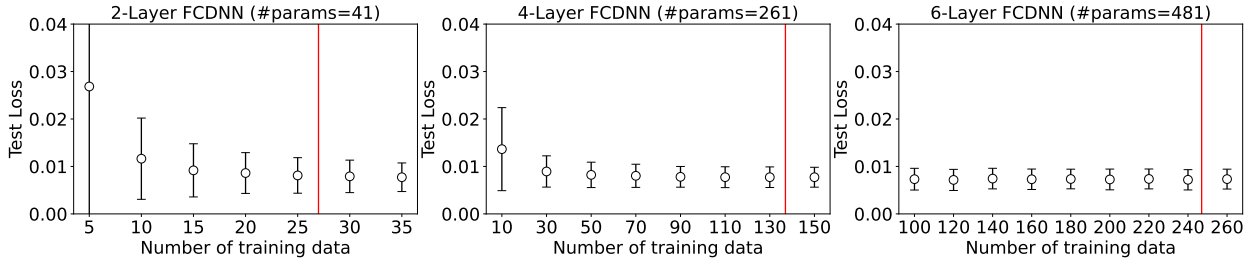

Figure 2: Test losses of random near interpolators for 2-layer (left), 4-layer (middle), and 6-layer (right) FCDNNs. The horizontal axis represents the number of training samples, and the vertical axis represents the test loss. Error bars indicate the standard deviation over 1000 sampled near interpolators for each training sample size. The red vertical line denotes the theoretical upper bound on the strong sample complexity from Theorem 4.

7 Conclusion

In this paper, we studied the generalization behavior of randomly sampled interpolators in noiseless teacher-student regression. We showed that the generalization error of a random interpolator becomes exactly zero once the number of training samples exceeds a threshold determined by the dimension of the teacher-equivalent set (TES).

Our analysis further showed that, for several deep neural network architectures, over-parameterization can substantially increase the TES dimension and thereby reduce the effective strong sample complexity. In particular, the strong sample complexity can remain much smaller than the total parameter count, which provides a partial theoretical explanation for why DNNs can generalize well in the over-parameterized regime, where the total number of parameters far exceeds the number of training samples.

Several important directions remain for future work. First, extending the theory to noisy teacher-student settings would provide a more realistic understanding of interpolation in practical machine learning. Second, our current framework relies heavily on analyticity assumptions and therefore does not directly apply to widely used non-analytic architectures such as ReLU networks. Extending the geometric analysis developed in this paper to non-smooth or piecewise-analytic models is an important open problem.

More broadly, our results suggest that the geometry of the teacher-equivalent set may provide a useful perspective for understanding why over-parameterized models can generalize despite possessing extremely large number of parameters.

References

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jonathan Bac, Evgeny M. Mirkes, Alexander N. Gorban, Ivan Tyukin, and Andrei Zinovyev. Scikit-dimension: A python package for intrinsic dimension estimation. *Entropy*, 23(10):1368, 2021.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

Mikhail Belkin, Daniel Hsu, and Partha P. Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Gon Buzaglo, Itamar Harel, Mor Shpigel Nacson, Alon Brutzkus, Nathan Srebro, and Daniel Soudry. How uniform random weights induce non-uniform bias: Typical interpolating neural networks generalize with narrow teachers. In *Proceedings of the International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Pingyeh Chiang, Renkun Ni, David Yu Miller, Arpit Bansal, Jonas Geiping, Micah Goldblum, and Tom Goldstein. Loss landscapes are all you need: Neural network generalization can be explained without the implicit bias of gradient descent. In *International Conference on Learning Representations*, 2023.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Proceedings of the Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, 2020.

Yaim Cooper. Global minima of overparameterized neural networks. *SIAM Journal on Mathematics of Data Science*, 3(2):676–691, 2021.

Assaf Dauber, Meir Feder, Tomer Koren, and Roi Livni. Can implicit bias explain generalization? stochastic convex optimization as a case study. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Alexander R. Farhang, Jeremy Bernstein, Kushal Tirumala, Yang Liu, and Yisong Yue. Investigating generalization by controlling normalized margin. In *Proceedings of the International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2022.

Herbert Federer. *Geometric Measure Theory*. Classics in Mathematics. Springer, 1996.

Charles Fefferman. Reconstructing a neural net from its output. *Revista Matemática Iberoamericana*, 10(3): 507–555, 1994.

Kenji Fukumizu, Shoichiro Yamaguchi, Yoh-ichi Mototake, and Mirai Tanaka. Semi-flat minima and saddle points by embedding neural networks to overparameterization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Keinosuke Fukunaga and David R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, C-20:176–183, 1971.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, 2010.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Robert M. Hardt. Stratification of real analytic mappings and images. *Inventiones Mathematicae*, 28(2): 193–208, 1975.

Robin Hartshorne. *Algebraic Geometry*. Graduate Texts in Mathematics. Springer, 1977.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel 'ridgeless' regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Proceedings of the Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, 2020.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75:667–766, 2022.

Chris Mingard, Guillermo Valle-Pérez, Joar Skalse, and Ard A. Louis. Is sgd a bayesian sampler? well, almost. *Journal of Machine Learning Research*, 22(1):3579–3642, 2021.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1:67–83, 2020.

Raghavan Narasimhan. *Introduction to the Theory of Analytic Spaces*, volume 25 of *Lecture Notes in Mathematics*. Springer, 1966.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *arXiv preprint arXiv:1911.00068*, 2020.

Kenta Oono and Taiji Suzuki. Approximation and non-parametric estimation of resnet-type convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2019.

Joel W. Robbin and Dietmar A. Salamon. Introduction to differential topology. Unpublished lecture notes, ETH Zürich, 2025.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *Proceedings of the International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2018.

Vahid Shahverdi, Giovanni Luca Marchetti, and Kathlén Kohn. Learning on a razor's edge: Identifiability and singularity of polynomial neural networks. In *International Conference on Learning Representations*, 2026.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, and Nathan Srebro. The implicit bias of gradient descent on separable data. In *International Conference on Learning Representations*, 2018.

Héctor J Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, 5(4):589–593, 1992.

Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In *Proceedings of the International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 2017.

Loring W. Tu. *An Introduction to Manifolds*. Universitext. Springer, 2008.

Konstantin Usevich, Ricardo Borsoi, Clara Dérand, and Marianne Clausel. Identifiability of deep polynomial neural networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Guillermo Valle-Pérez, Chico Q. Camargo, and Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *International Conference on Learning Representations*, 2019.

Verner Vlačić and Helmut Bölcskei. Neural network identifiability for a family of sigmoidal nonlinearities. *Constructive Approximation*, 55:173–224, 2022.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Gal Kaplun, Sham M. Kakade, and Boaz Barak. Beyond implicit bias: The insignificance of sgd noise in online learning. In *Proceedings of the International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2009.

Hassler Whitney and François Bruhat. Quelques propriétés fondamentales des ensembles analytiques-réels. *Commentarii Mathematici Helvetici*, 33:132–160, 1959.

Weihang Xu and Simon S. Du. Over-parameterization exponentially slows down gradient descent for learning a single neuron. In *Proceedings of the Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, 2023.

Zitong Yang, Yu Bai, and Song Mei. Exact gap between generalization error and uniform convergence in random feature models. In *Proceedings of the International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 2021.

Lijia Yu, Xiao-Shan Gao, Lijun Zhang, and Yibo Miao. Generalizability of memorization neural network. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A unifying view on implicit bias in training linear neural networks. In *International Conference on Learning Representations*, 2021.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

A Additional Information Omitted in the Main Text

A.1 Formal definition of the uniform distribution on $\widehat{\Theta}_n$ in Section 3.1

In this section, we make precise the meaning of the uniform distribution on the interpolator set $\widehat{\Theta}_n$.

Since $\widehat{\Theta}_n$ is typically a lower-dimensional subset of Θ, the standard Lebesgue measure on Θ restricted to $\widehat{\Theta}_n$ cannot be normalized into a probability measure. We therefore define the uniform distribution on $\widehat{\Theta}_n$ using the intrinsic $\dim(\widehat{\Theta}_n)$-dimensional volume measure on the top-dimensional regular part of $\widehat{\Theta}_n$.

Let $S \subset \Theta$ be a compact real analytic set, and let

$$r := \dim(S).$$

We define the r-dimensional regular part of S, denoted by $S_{\mathrm{reg}}^{(r)}$, as the set of points $\theta \in S$ for which there exists an open neighborhood $U \subset \Theta$ of θ such that $S \cap U$ is a real analytic submanifold of Θ of dimension r. Equivalently, around every point of $S_{\mathrm{reg}}^{(r)}$, there exists a real analytic parametrization

$$\psi : V \subset \mathbb{R}^r \to S \cap U$$

whose image is an open neighborhood of the point in $S_{\mathrm{reg}}^{(r)}$ and whose differential has rank r.

We define the intrinsic Lebesgue measure on S as the Riemannian volume measure on $S_{\mathrm{reg}}^{(r)}$ induced by the Euclidean metric of the parameter space $\Theta \subset \mathbb{R}^{d_\Theta}$. More explicitly, for a local analytic parametrization

$$\psi : V \subset \mathbb{R}^r \to S_{\mathrm{reg}}^{(r)},$$

the local volume element is given by

$$\mathrm{d}\lambda_S(\psi(u)) = \sqrt{\det\left(D\psi(u)^\top D\psi(u)\right)}\,\mathrm{d}u,$$

where $\sqrt{\det\left(D\psi(u)^\top D\psi(u)\right)}$ is the r-dimensional volume Jacobian of the local parametrization ψ and $\mathrm{d}u$ denotes the standard Lebesgue measure on $\mathbb{R}^r$. These local definitions are compatible under analytic changes of coordinates, and hence define a Borel measure λ_S on S, supported on $S_{\mathrm{reg}}^{(r)}$. We call λ_S the intrinsic Lebesgue measure on S.

Since S is compact and real analytic, $S_{\mathrm{reg}}^{(r)}$ has finite intrinsic volume. Moreover, if $S \neq \emptyset$, then

$$0 < \lambda_S(S) < \infty.$$

Therefore, the intrinsic uniform distribution on S is defined by

$$U_S(A) := \frac{\lambda_S(A \cap S_{\mathrm{reg}}^{(r)})}{\lambda_S(S)}$$

for every Borel set $A \subset S$.

For the interpolator set $\widehat{\Theta}_n$, we write $U_{\widehat{\Theta}_n}$ for this intrinsic uniform distribution. Throughout the paper, the random interpolator distribution is assumed to satisfy

$$\mathbb{P}(\cdot \mid \widehat{\Theta}_n) \ll U_{\widehat{\Theta}_n}.$$

That is, the distribution of the random interpolator is absolutely continuous with respect to the intrinsic uniform distribution on the top-dimensional regular part of $\widehat{\Theta}_n$.

B Basic Concepts of Real Analytic Sets

We introduce basic concepts and results in real analytic sets. Consistently with the main contents, let $\Theta \subset \mathbb{R}^{d_\Theta}$ be a compact real analytic manifold, possibly with boundary.

B.1 Irreducible sets

First, we see the property of irreducible sets. We define the irreducibility of a real analytic set as follows.

Definition 7 (Irreducibility)**.** A real analytic set X is called *irreducible* if it is not the union of two strictly smaller real analytic sets.

We show two important theorems about the irreducible sets.

Theorem 6 (Proposition 11, Whitney & Bruhat (1959))**.** *For a real analytic set X, there uniquely exists a locally finite family $\{S_\lambda\}_{\lambda\in\Lambda}$ of irreducible real analytic subsets of X such that $X = \bigcup_\lambda S_\lambda$ and $S_\lambda \not\subset S_\mu$ for $\lambda, \mu \in \Lambda$ with $\lambda \neq \mu$.*

Theorem 7 (Corollary, Section 8, Whitney & Bruhat (1959))**.** *An irreducible real analytic set contains no proper real analytic subset of its same dimension.*

From Theorem 6, we can show the following immediately.

Proposition 8. *A real analytic set X defined on a compact space Θ can be decomposed into a finite number of irreducible real analytic sets.*

Proof. From Theorem 6, there exists a locally finite family of irreducible real analytic sets $\{S_\lambda\}_\lambda$ such that $X = \bigcup_\lambda S_\lambda$ holds. Since X is a closed subset of Θ, X is also compact. Therefore, since $\{S_\lambda\}_\lambda$ is locally finite cover and X is compact, we can choose a finite number of subsets from $\{S_\lambda\}_\lambda$ covering X. These subsets are the desired irreducible real analytic sets. □

B.2 Regularity

Second, we introduce the notion of regularity. We consider a real analytic set $X \subset \Theta$ and define $d := \dim(X)$.

Definition 8 (Analytic isomorphism)**.** Two sets $X_1, X_2 \subset \mathbb{R}^d$ are said to be analytically isomorphic if there exists an analytic isomorphism from X_1 to X_2.

We define the regularity of a point in a real analytic set as follows. This definition follows Narasimhan (1966).

Definition 9 (Regularity of a point)**.** For a fixed point $x \in X$, x is said to be regular if there exists an open neighborhood $U \subset \Theta$ of x such that $X \cap U$ is analytically isomorphic to some open set V in $\mathbb{R}^d$.

As an important property of regular points, we show the following theorem.

Theorem 9 (Theorem 1, Chapter III, Narasimhan (1966))**.** *The set of regular points in X is dense in X.*

C Proof of Theorem 2

Proof. First, we show that, regardless of the choice of $x_1, ..., x_n$, the dimension of $\widehat{\Theta}_{n+1} \setminus \bar{\Theta}$ becomes less than that of $\widehat{\Theta}_n \setminus \bar{\Theta}$ with probability 1, with respect to the randomness of x_{n+1}. Here, the dimension of the real analytic set difference $\widehat{\Theta}_n \setminus \bar{\Theta}$ is defined by the maximum dimension of real analytic manifolds included in $\widehat{\Theta}_n \setminus \bar{\Theta}$. Second, we show that the dimension of $\widehat{\Theta}_n \setminus \bar{\Theta}$ becomes at least $n-1$ less than that of $\widehat{\Theta}_1 \setminus \bar{\Theta}$ with probability 1, with respect to the randomness of $x_1, ..., x_n$. Finally, we complete the proof by characterizing the properties of $\widehat{\theta}_n$ sampled from $\widehat{\Theta}_n$.

Step 0: The dimension of $\widehat{\Theta}_1 \setminus \bar{\Theta}$ is less than d_Θ with probability 1. Before proceeding to the main part of the proof, we first show that the dimension of $\widehat{\Theta}_1 \setminus \bar{\Theta}$ is less than d_Θ as a base case.

For $x \in \mathcal{X}$, define

$$G_x(\theta) := \ell\left(f^*(x;\theta^*), f(x;\theta)\right).$$

Let $C_1, \ldots, C_M$ be the connected components of the compact real analytic manifold Θ. Since Θ is compact, the number of connected components is finite.

We show that

$$\dim(\widehat{\Theta}_1 \setminus \bar{\Theta}) \leq d_\Theta - 1$$

with probability one over $x_1 \sim D$.

Fix a connected component C_j. If $C_j \subset \bar{\Theta}$, then

$$(C_j \cap \widehat{\Theta}_1) \setminus \bar{\Theta} = \emptyset,$$

so this component contributes nothing to $\widehat{\Theta}_1 \setminus \bar{\Theta}$.

Now suppose $C_j \not\subset \bar{\Theta}$. Choose

$$\theta_j \in C_j \setminus \bar{\Theta}.$$

Define

$$E_j := \{x \in \mathcal{X} : G_x|_{C_j} \equiv 0\}.$$

Then

$$E_j \subset \{x \in \mathcal{X} : G_x(\theta_j) = 0\}.$$

Since $\theta_j \notin \bar{\Theta}$, the analytic function

$$x \mapsto G_x(\theta_j)$$

is not identically zero on $\mathcal{X}$. Therefore, by the identity theorem for real analytic functions, its zero set is a proper analytic subset of $\mathcal{X}$, and hence has measure zero with respect to the uniform measure on $\mathcal{X}$. Hence, when we sample $x_1 \sim \mathcal{D}$, we have

$$x_1 \notin E_j$$

with probability 1 from Assumption 1.

Since there are only finitely many connected components C_j, with probability 1 over $x_1 \sim \mathcal{D}$, we have

$$x_1 \notin \bigcup_{j : C_j \not\subset \bar{\Theta}} E_j.$$

For such an x_1, the analytic function $G_{x_1}|_{C_j}$ is not identically zero on every connected component $C_j \not\subset \bar{\Theta}$. Hence

$$C_j \cap \widehat{\Theta}_1 = \{\theta \in C_j : G_{x_1}(\theta) = 0\}$$

is a proper analytic subset of the d_Θ-dimensional analytic manifold C_j. Therefore,

$$\dim(C_j \cap \widehat{\Theta}_1) \leq d_\Theta - 1.$$

Combining this over all connected components, we obtain

$$\dim(\widehat{\Theta}_1 \setminus \bar{\Theta}) \leq d_\Theta - 1$$

with probability one.

Step 1: The dimension of $\widehat{\Theta}_{n+1} \setminus \bar{\Theta}$ is less than that of $\widehat{\Theta}_n \setminus \bar{\Theta}$ with probability 1. We prove this by using the knowledge of analytic sets as we see in section B. As we show in Proposition 8, $\widehat{\Theta}_n$ $(n \geq 1)$ can be decomposed into a finite number of irreducible real analytic sets. We define a sequence of irreducible real analytic sets that are not fully included in $\bar{\Theta}$ as

$$T_{n,1}, ..., T_{n,k} \subset \widehat{\Theta}_n.$$

If there does not exist such a set, $\bar{\Theta} = \widehat{\Theta}_n$ holds, and the proof is completed, so we assume the existence of such a sequence. Our goal is to show that the dimension of $T_{n,i} \cap \{\theta \in \Theta \mid \ell(f^*(x_{n+1}; \theta^*), f(x_{n+1}; \theta)) = 0\}$ is less than that of $T_{n,i}$ with probability 1 in terms of the randomly chosen x_{n+1} for all $i = 1, \cdots, k$. Choose

one arbitrarily from the sequence of sets and denote it by T. Let the dimension of T be d_T. d_T may vary depending on the choice of $T_{n,i}$, but it does not affect the following proof.

We prepare some useful notions. We fix an arbitrary point $\theta' \in T$ and define

$$\mathcal{X}_{\theta'} := \{x \in \mathcal{X} \mid \ell(f^*(x;\theta^*), f(x;\theta')) = 0\}.$$

Furthermore, we denote

$$\mathcal{X}_T := \bigcap_{\theta' \in T} \mathcal{X}_{\theta'} = \{x \in \mathcal{X} | \{\theta \in \widehat{\Theta}_n \mid \ell(f^*(x;\theta^*), f(x;\theta)) = 0\} \supset T\}.$$

From the definition, we have $\mathcal{X}_T \subset \mathcal{X}_{\theta'}$. In addition, $\mathcal{X}_{\theta'}$ is a real analytic set with respect to x.

We show that the dimension of $\mathcal{X}_T$ is less than m, which is the dimension of $\mathcal{X}$, and its Lebesgue measure on $\mathcal{X}$ is equal to 0. To this aim, we assume that the dimension of $\mathcal{X}_T$ is m and prove that it is a contradiction. From this assumption, the dimension of $\mathcal{X}_{\theta'}$ is also m. As we see in Theorem 9, the set of regular points in $\mathcal{X}_{\theta'}$ is dense. Hence, we can choose one of the regular points of $\mathcal{X}_{\theta'}$ and its open neighborhood $U \subset \mathcal{X}$. From the regularity, $\mathcal{X}_{\theta'} \cap U$ can be mapped into an open set $\widehat{\mathcal{X}}_{\theta'}$ in $\mathbb{R}^m$ by an analytic isomorphism since the dimension of $\mathcal{X}_{\theta'}$ is m. That is, we have $g_{\mathcal{X}}(\mathcal{X}_{\theta'} \cap U) = \widehat{\mathcal{X}}_{\theta'}$ for an analytic isomorphism $g_{\mathcal{X}} : U \to V$ where V is an open set in $\mathbb{R}^m$. Hence, we have

$$\ell(f^*(x;\theta^*), f(x;\theta')) = 0, \forall x \in g_{\mathcal{X}}^{-1}(\widehat{\mathcal{X}}_{\theta'}).$$

Since $\widehat{\mathcal{X}}_{\theta'}$ is a nonempty open subset of V and $g_{\mathcal{X}} : U \to V$ is a homeomorphism, $g_{\mathcal{X}}^{-1}(\widehat{\mathcal{X}}_{\theta'})$ is a nonempty open subset of U, and hence of $\mathcal{X}$. Therefore, the identity theorem shows that $\ell(f^*(x;\theta^*), f(x;\theta'))$ is equal to 0 for every $x \in \mathcal{X}$ since $\mathcal{X}$ is connected. Since the choice of θ' is arbitrary, we have

$$\ell(f^*(x;\theta^*), f(x;\theta')) = 0, \forall x \in \mathcal{X}, \forall \theta' \in T.$$

However, it means that T is included in $\bar{\Theta}$, so it contradicts the assumption to the definition of T. Hence, the dimension of $\mathcal{X}_T$ is less than m. So the Lebesgue measure of $\mathcal{X}_T$ on $\mathcal{X}$ is 0.

Let x_{n+1} be a random sample from the distribution $\mathcal{D}$. Since $\mathcal{D}$ is absolutely continuous to a uniform distribution on $\mathcal{X}$, we have from the discussion above,

$$\{\theta \in \widehat{\Theta}_n \mid \ell(f^*(x_{n+1};\theta^*), f(x_{n+1};\theta)) = 0\} \cap T_{n,i} \subsetneq T_{n,i} \tag{3}$$

with probability 1 for every $i = 1, ..., k$.

Let $S_{n+1} := \{\theta \in \widehat{\Theta}_n \mid \ell(f^*(x_{n+1};\theta^*), f(x_{n+1};\theta)) = 0\}$. Since S_{n+1} is a real analytic set and $T_{n,i}$ is irreducible, (3) means that the dimension of $S_{n+1} \cap T_{n,i}$ is less than that of $T_{n,i}$ from Theorem 7. Hence, the dimension of $S_{n+1} \cap T_{n,i}$ is less than that of $T_{n,i}$ with probability 1 for all the components $T_{n,i}$ of $\widehat{\Theta}_n$. Note that if the dimension of $T_{n,i}$ is 0, $S_{n+1} \cap T_{n,i}$ is an empty set. In addition, we have

$$\begin{aligned}\widehat{\Theta}_{n+1} \setminus \bar{\Theta} =& (\widehat{\Theta}_n \cap S_{n+1}) \setminus \bar{\Theta} = \bigcup_{i=1}^{k} (T_{n,i} \cap S_{n+1}) \setminus \bar{\Theta} \\ \widehat{\Theta}_n \setminus \bar{\Theta} =& \bigcup_{i=1}^{k} T_{n,i} \setminus \bar{\Theta}.\end{aligned}$$

Therefore, the dimension of $\widehat{\Theta}_{n+1} \setminus \bar{\Theta}$ is less than that of $\widehat{\Theta}_n \setminus \bar{\Theta}$ with probability 1. Note that when the dimension of $\widehat{\Theta}_n \setminus \bar{\Theta}$ is 0, $\widehat{\Theta}_{n+1} \setminus \bar{\Theta}$ is an empty set: since $T_{n,i}$ is of dimension 0 , $S_{n+1} \cap T_{n,i}$ is an empty set for $i \in [k]$.

Step 2: The dimension of $\widehat{\Theta}_n \setminus \bar{\Theta}$ is at least n less than d_Θ with probability 1. We denote by $\widehat{\Theta}(x_1, \ldots, x_n)$ the zero set $\widehat{\Theta}_n$ determined by $x_1, \ldots, x_n$. We also define the event that the dimension of

$\widehat{\Theta}(x_1, \ldots, x_n, x_{n+1}) \setminus \bar{\Theta}$ is less than that of $\widehat{\Theta}(x_1, \ldots, x_n) \setminus \bar{\Theta}$ by $A(x_1, ..., x_{n+1})$. $A(x_1)$ denotes the event that the dimension of $\widehat{\Theta}(x_1) \setminus \bar{\Theta}$ is less than d_Θ. What we aim to show is that, when $x_1, \ldots, x_n$ are i.i.d. generated random variables from $\mathcal{D}$,

$$\mathbb{P}(A(x_1, ..., x_n) \cap A(x_1, ..., x_{n-1}) \cap \cdots \cap A(x_1)) = 1.$$

From Step 0, the probability of a random variable $x_1 \sim \mathcal{D}$ causing an event $A(x_1)$ is given by

$$\mathbb{P}(A(x_1)) = 1. \tag{4}$$

Next, observe that for i.i.d. random variables $x_1, x_2 \sim \mathcal{D}$, we have

$$\begin{aligned}\mathbb{P}\left(\overline{A(x_1, x_2) \cap A(x_1)}\right) =& \mathbb{P}\left(\overline{A(x_1, x_2)} \cup \overline{A(x_1)}\right) \\ \leq& \mathbb{P}\left(\overline{A(x_1, x_2)}\right) + \mathbb{P}\left(\overline{A(x_1)}\right) \\ =& \mathbb{P}\left(\overline{A(x_1, x_2)}\right)\end{aligned} \tag{5}$$

holds from (4). Moreover, we have

$$\mathbb{P}\left(\overline{A(x_1, x_2)}\right) = \int_{x_1' \in \mathcal{X}} \mathbb{P}\left(\overline{A(x_1', x_2)} \mid x_1'\right) \cdot \mathrm{d}\mathbb{P}(x_1'),$$

where $\mathbb{P}\left(\overline{A(x_1', x_2)} \mid x_1'\right)$ denotes a probability measure of the event that $A(x_1', x_2)$ does not occur given that $x_1 = x_1'$, and $\mathrm{d}\mathbb{P}(x_1')$ denotes the probability measure of the event that $x_1 = x_1'$ occurs for fixed $x_1' \in \mathcal{X}$. Since $\mathbb{P}(A(x_1', x_2) \mid x_1') = 1$ holds for any $x_1' \in \mathcal{X}$ from Step 1, we have

$$\int_{x_1' \in \mathcal{X}} \mathbb{P}\left(\overline{A(x_1', x_2)} \mid x_1'\right) \cdot \mathrm{d}\mathbb{P}(x_1') = 0 \cdot 1 = 0.$$

Therefore, we have from (5),

$$\mathbb{P}(A(x_1, x_2) \cap A(x_1)) = 1.$$

Repeating this argument inductively, we obtain

$$\mathbb{P}(A(x_1, ..., x_n) \cap A(x_1, ..., x_{n-1}) \cap \cdots \cap A(x_1)) = 1$$

as desired.

Step 3: The properties of $\widehat{\theta}_n$ sampled from $\widehat{\Theta}_n$. From Step 0 to Step 2, when $n \geq d_\Theta - d_{\bar{\Theta}} + 1$, the dimension of $\widehat{\Theta}_n \setminus \bar{\Theta}$ is less than that of $\bar{\Theta}$ and therefore, less than that of $\widehat{\Theta}_n$. Note that if the dimension of $\bar{\Theta}$ is 0, $\widehat{\Theta}_n \setminus \bar{\Theta}$ is an empty set when $n \geq d_\Theta + 1$. As a consequence, the intrinsic Lebesgue measure of $\widehat{\Theta}_n \setminus \bar{\Theta}$ on $\widehat{\Theta}_n$ defined in Section A.1 corresponds to 0. Since $\mathbb{P}(\cdot|\widehat{\Theta}_n)$ is absolutely continuous to the intrinsic uniform distribution on $\widehat{\Theta}_n$, we have for $\widehat{\theta}_n \sim \mathbb{P}(\cdot|\widehat{\Theta}_n)$

$$\widehat{\theta}_n \in \bar{\Theta}$$

with probability 1, which completes the proof. □

D Proof of Proposition 3

D.1 The asymptotic form of the volume of a neighborhood

In the proof, we utilize a theory developed by Federer (1996) with regard to the asymptotic form of the volume of an ε-neighborhood of a manifold. For completeness of the paper, we present basic concepts of this theory in this section.

First, we present a notion of Minkowski content, which defines an asymptotic form of the volume of an ε-neighborhood for small ε.

Definition 10 (Minkowski content)**.** Let A be a Lebesgue measurable set in $\mathbb{R}^n$ and $\mathrm{Vol}(A)$ be its Lebesgue measure on $\mathbb{R}^n$. If there exists the following value for $S \subset \mathbb{R}^n$, we call it m-dimensional Minkowski content and denote it as $\mathcal{M}^m(S)$:

$$\lim_{\varepsilon \to 0+} \frac{\mathrm{Vol}(\{x \mid \|x - S\| \leq \varepsilon\})}{\alpha(n-m)\varepsilon^{n-m}},$$

where $\alpha(m)$ is the Lebesgue measure of a unit ball in $\mathbb{R}^m$.

Definition 11 (Rectifiable)**.** Let S be a subset of a metric space X and m be a positive integer. Then S is called *m-rectifiable* if and only if there exists a Lipschitzian function mapping some bounded subset of $\mathbb{R}^m$ onto S.

From the definition, a compact m-dimensional smooth manifold in $\mathbb{R}^n$ is m-rectifiable.

Definition 12 (Hausdorff measure)**.** Let X be a metric space with distance ρ and S be a Carathéodory-measurable set in it. We consider a δ-covering $\{U_i^\delta\}$ of S as

$$S \subset \bigcup_{i=1}^{\infty} U_i^\delta, \mathrm{diam}(U_i^\delta) \leq \delta,$$

where $\mathrm{diam}(U_i^\delta) := \sup_{x,y \in U_i^\delta} \rho(x,y)$. We define

$$\mathcal{H}_\delta^m(S) := \inf_{\{U_i^\delta\}} \sum_{i=1}^{\infty} \mathrm{diam}(U_i^\delta)^m.$$

Then we call the following value *m-dimensional Hausdorff measure* of S.

$$\mathcal{H}^m(S) := \lim_{\delta \to 0} \mathcal{H}_\delta^m(S).$$

We note that while the Hausdorff measure is not necessarily a finite value, when S is an m-dimensional smooth compact manifold in $\mathbb{R}^n$, the m-dimensional Hausdorff measure of it becomes a positive finite value (see section 3.2.46 in Federer (1996) for instance).

We now present the main result necessary for the proof of Proposition 3.

Theorem 10 (Theorem 3.2.39 (Federer, 1996))**.** *If S is a closed m-rectifiable subset of $\mathbb{R}^n$, then we have*

$$\mathcal{M}^m(S) = \mathcal{H}^m(S).$$

From this theorem, the following holds immediately.

Corollary 11. *For small ε, we have an asymptotic form for the Lebesgue measure of an ε-neighborhood of an m-dimensional smooth compact manifold with boundary S in $\mathbb{R}^n$ as*

$$\mathrm{Vol}(\{x \mid \|x - S\| \leq \varepsilon\}) = \alpha(n-m)\varepsilon^{n-m}\mathcal{H}^m(S) + o(\varepsilon^{n-m}).$$

D.2 Proof of Proposition 3

Proof. From now, we denote the ε-neighborhood of a set $A \subset \Theta$ as

$$A_\varepsilon := \{\theta \in \Theta \mid \|\theta - A\| \leq \varepsilon\}$$

and the Lebesgue measure of A on Θ as $\mathrm{Vol}(A)$. We also denote the dimension of A as $\dim(A)$. We prove this theorem by showing the following two steps:

1. The sampled $\widehat{\theta}_{n,\varepsilon}$ lies in the ε-neighborhood of $\bar{\Theta}$ with high probability at least $1 - c\varepsilon$ for some constant c.

2. If $\widehat{\theta}_{n,\varepsilon}$ is sampled from this neighborhood, then the generalization error becomes no more than $(q\varepsilon)^2$.

First, we evaluate the probability of sampling $\widehat{\theta}_{n,\varepsilon}$ from the ε-neighborhood of $\bar{\Theta}$. Since we sample $\widehat{\theta}_{n,\varepsilon}$ from $\widehat{\Theta}_{n,\varepsilon}$ uniformly, this probability is written as

$$\frac{\mathrm{Vol}(\bar{\Theta}_\varepsilon)}{\mathrm{Vol}(\widehat{\Theta}_{n,\varepsilon})}.$$

We can decompose the term as

$$\mathrm{Vol}(\widehat{\Theta}_{n,\varepsilon}) \leq \mathrm{Vol}(\bar{\Theta}_\varepsilon) + \mathrm{Vol}((\widehat{\Theta}_n \setminus \bar{\Theta})_\varepsilon)$$

since $\widehat{\Theta}_{n,\varepsilon} \subset \bar{\Theta}_\varepsilon \cup (\widehat{\Theta}_n \setminus \bar{\Theta})_\varepsilon$ naturally holds. Hence we have

$$\begin{aligned}\frac{\mathrm{Vol}(\bar{\Theta}_\varepsilon)}{\mathrm{Vol}(\widehat{\Theta}_{n,\varepsilon})} \geq& \frac{\mathrm{Vol}(\bar{\Theta}_\varepsilon)}{\mathrm{Vol}(\bar{\Theta}_\varepsilon) + \mathrm{Vol}((\widehat{\Theta}_n \setminus \bar{\Theta})_\varepsilon)} \\ =& 1 - \frac{\mathrm{Vol}((\widehat{\Theta}_n \setminus \bar{\Theta})_\varepsilon)}{\mathrm{Vol}(\bar{\Theta}_\varepsilon) + \mathrm{Vol}((\widehat{\Theta}_n \setminus \bar{\Theta})_\varepsilon)} \\ \geq& 1 - \frac{\mathrm{Vol}((\widehat{\Theta}_n \setminus \bar{\Theta})_\varepsilon)}{\mathrm{Vol}(\bar{\Theta}_\varepsilon)}. \end{aligned} \tag{6}$$

Since $\bar{\Theta}$ contains a $\dim(\bar{\Theta})$-dimensional analytic manifold, Corollary 11 shows that there exists a positive constant c_1 such that

$$\mathrm{Vol}(\bar{\Theta}_\varepsilon) \geq c_1 \varepsilon^{\dim(\Theta)-\dim(\bar{\Theta})} + o(\varepsilon^{\dim(\Theta)-\dim(\bar{\Theta})}). \tag{7}$$

Next, observe that $\widehat{\Theta}_n \setminus \bar{\Theta}$ is a semi-analytic set, which is a parameter set defined by equations and inequalities of analytic functions. By Section 3 in Hardt (1975), a semi-analytic set defined on a compact parameter set is decomposed into a finite number of smooth manifolds. Hence, if we decompose $\widehat{\Theta}_n \setminus \bar{\Theta}$ into smooth manifolds $\widehat{\Theta}_n \setminus \bar{\Theta} = Z^1 \cup \cdots \cup Z^k$, we have

$$\mathrm{Vol}((\widehat{\Theta}_n \setminus \bar{\Theta})_\varepsilon) \leq \sum_{i=1}^{k} \mathrm{Vol}(Z^i_\varepsilon).$$

Since each Z^i has a dimension of no more than $\dim(\widehat{\Theta}_n \setminus \bar{\Theta})$, we have from Corollary 11,

$$\mathrm{Vol}((\widehat{\Theta}_n \setminus \bar{\Theta})_\varepsilon) \leq c_2 \varepsilon^{\dim(\Theta)-\dim(\widehat{\Theta}_n \setminus \bar{\Theta})} + o(\varepsilon^{\dim(\Theta)-\dim(\widehat{\Theta}_n \setminus \bar{\Theta})}) \tag{8}$$

for a positive constant c_2.

From the proof of Theorem 2, we have

$$\dim(\widehat{\Theta}_n \setminus \bar{\Theta}) \leq \dim(\bar{\Theta}) - 1 \tag{9}$$

almost surely with respect to the training samples.

Substituting (7), (8) and (9) into (6), we have

$$\mathrm{Vol}(\bar{\Theta}_\varepsilon)/\mathrm{Vol}(\widehat{\Theta}_{n,\varepsilon}) \geq 1 - c\varepsilon$$

for a finite positive constant c depending on the training samples. From the discussion above, $\widehat{\theta}_{n,\varepsilon}$ sampled from $\widehat{\Theta}_{n,\varepsilon}$ is in $\bar{\Theta}_\varepsilon$ with probability at least $1 - c\varepsilon$.

Second, we evaluate the generalization error when we sample $\widehat{\theta}_{n,\varepsilon}$ from $\bar{\Theta}_\varepsilon$. Since $f(x;\theta)$ is q-Lipschitz continuous with respect to θ for every $x \in \mathcal{X}$, we have for some $\bar{\theta} \in \bar{\Theta}$,

$$\begin{aligned}
L(\widehat{\theta}_{n,\varepsilon}) =& \mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{2}\left\|f(x;\widehat{\theta}_{n,\varepsilon}) - f^*(x;\theta^*)\right\|^2\right] \\
=& \mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{2}\left\|(f(x;\widehat{\theta}_{n,\varepsilon}) - f(x;\bar{\theta})) + (f(x;\bar{\theta}) - f^*(x;\theta^*))\right\|^2\right] \\
\leq& \mathbb{E}_{x\sim\mathcal{D}}\left[\left\|f(x;\widehat{\theta}_{n,\varepsilon}) - f(x;\bar{\theta})\right\|^2\right] + \mathbb{E}_{x\sim\mathcal{D}}\left[\left\|f(x;\bar{\theta}) - f^*(x;\theta^*)\right\|^2\right] \\
\leq& (q\|\widehat{\theta}_{n,\varepsilon} - \bar{\theta}\|)^2 + 0 \\
\leq& (q\varepsilon)^2,
\end{aligned}$$

which completes the proof. □

E Proof of Theorem 4

The loss function of an FCDNN satisfies Assumption 2 since linear transformations and activations by an analytic function are both analytic.

Therefore, by Theorem 2, it suffices to study the lower bound on $d_{\bar{\Theta}}$. We proceed in the following steps:

1. Construct a subset $\bar{\Theta}'$ of the TES $\bar{\Theta}$.
2. Count the number of free parameters of $\bar{\Theta}'$, which gives a lower bound on $d_{\bar{\Theta}}$.

First, we define the following subset $\bar{\Theta}'$ of Θ.

Definition 13 (Subset $\bar{\Theta}'$ of Θ)**.** Let us denote arbitrary matrices or vectors as $\mathcal{M}$. $\bar{\Theta}'$ is the subset of Θ which satisfies the following condition. Superscripts represent the layer index and subscripts are for notational distinction.

1. For the parameter of first layer,

$$w^{(1)} = \begin{pmatrix} {w^{(1)}}^* \\ \mathcal{M}_1^{(1)} \end{pmatrix}, b^{(1)} = \begin{pmatrix} {b^{(1)}}^* \\ \mathcal{M}_2^{(1)} \end{pmatrix}.$$

2. For the parameter of ℓ-th layer ($2 \leq \ell \leq L-1$),

$$w^{(\ell)} = \begin{pmatrix} {w^{(\ell)}}^* & 0 \\ \mathcal{M}_1^{(\ell)} & \mathcal{M}_2^{(\ell)} \end{pmatrix}, b^{(\ell)} = \begin{pmatrix} {b^{(\ell)}}^* \\ \mathcal{M}_3^{(\ell)} \end{pmatrix}.$$

3. For the parameter of L-th layer,

$$w^{(L)} = \begin{pmatrix} {w^{(L^*)}}^* & 0 \end{pmatrix}, b^{(L)} = {b^{(L^*)}}^*.$$

We show the following.

Lemma 12. *$\bar{\Theta}'$ is a subset of $\bar{\Theta}$.*

Proof. We show that the student network whose parameters are in $\bar{\Theta}'$ exactly replicates the teacher network. We denote the pre-activation in the ℓ-th layer of the teacher network as ${h^{(\ell)}}^*$ ($\in \mathbb{R}^{m_\ell^*}$) and the pre-activation in the redundant width of the student network in the ℓ-th layer as $r^{(\ell)}$.

The pre-activation in the first layer of the student network is

$$\begin{aligned} w^{(1)}x + b^{(1)} &= \begin{pmatrix} {w^{(1)}}^* x \\ \mathcal{M}_1^{(1)} x \end{pmatrix} + \begin{pmatrix} {b^{(1)}}^* \\ \mathcal{M}_2^{(1)} \end{pmatrix} \\ &= \begin{pmatrix} {h^{(1)}}^* \\ r^{(1)} \end{pmatrix}. \end{aligned}$$

The pre-activation in the second layer of the student network is

$$\begin{aligned} w^{(2)}\sigma\left(\begin{pmatrix} {h^{(1)}}^* \\ r^{(1)} \end{pmatrix}\right) + b^{(2)} &= \begin{pmatrix} {w^{(2)}}^* \sigma({h^{(1)}}^*) + 0\sigma(r^{(1)}) \\ \mathcal{M}_1^{(2)}\sigma({h^{(1)}}^*) + \mathcal{M}_2^{(2)}\sigma(r^{(1)}) \end{pmatrix} + \begin{pmatrix} {b^{(2)}}^* \\ \mathcal{M}_3^{(2)} \end{pmatrix} \\ &= \begin{pmatrix} {h^{(2)}}^* \\ r^{(2)} \end{pmatrix}. \end{aligned}$$

Repeating this argument for L layers, the pre-activation of $(L-1)$-th layer can be written as $\begin{pmatrix} {h^{(L-1)}}^* \\ r^{(L-1)} \end{pmatrix}$.

Hence, the output of the last layer is

$${w^{(L)}}^*\sigma({h^{(L-1)}}^*) + 0\sigma(r^{(L-1)}) + {b^{(L)}}^* = {w^{(L)}}^*\sigma({h^{(L-1)}}^*) + {b^{(L)}}^*,$$

which is the same as the output from the teacher model. □

Next, we obtain the lower bound for the dimension of $\bar{\Theta}$ by counting the number of constrained parameters in $\bar{\Theta}'$.

Lemma 13. *We have*

$$d_{\bar{\Theta}} \geq d_{\Theta} - d',$$

where

$$d' = m_1^*(m+1) + \sum_{\ell=2}^{L-1}(m_\ell^*(m_{\ell-1}+1)) + p(m_{L-1}+1).$$

Proof. From Lemma 12, $\bar{\Theta}$ contains $\bar{\Theta}'$. The number of constrained parameters in $\bar{\Theta}'$ is

$$\begin{aligned} d' :=& m_1^* m + m_1^* + \sum_{\ell=2}^{L-1}(m_\ell^* m_{\ell-1} + m_\ell^*) + p m_{L-1} + p \\ =& m_1^*(m+1) + \sum_{\ell=2}^{L-1}(m_\ell^*(m_{\ell-1}+1)) + p(m_{L-1}+1). \end{aligned}$$

Hence, $\bar{\Theta}'$ is a $(d_\Theta - d')$-dimensional hypercube, and $\bar{\Theta}$ contains it. Since a $(d_\Theta - d')$-dimensional hypercube contains an analytic manifold of dimension $d_\Theta - d'$, the dimension of $\bar{\Theta}$ is lower bounded by $d_\Theta - d'$. □

Substituting the result of Lemma 13 into Theorem 2, we obtain $\widehat{k} \leq d' + 1$, which completes the proof of Theorem 4.

F Proof of Theorem 5

All the operations in a ResNet defined in Section 5.2 are a linear transformation or an activation mapping by an analytic function, therefore, analytic. Hence, Assumption 2 is satisfied.

Let the student parameters in the ℓ-th layer of the i-th residual block be $(w^{(i,\ell)}, b^{(i,\ell)})$ $(i = 1, \cdots, N$, $\ell = 1, \cdots, L)$ and those in the fully connected layer be (w, b). Let the teacher parameters in the ℓ-th layer

of the i-th residual block be $(w^{(i,\ell)^*}, b^{(i,\ell)^*})$ $(i = 1, \cdots, N^*,\ \ell = 1, \cdots, L)$ and those in the fully connected layer be (w^*, b^*).

We define the following subset $\bar{\Theta}'$ of Θ.

Definition 14 (Subset $\bar{\Theta}'$ of Θ)**.** Let arbitrary matrices or vectors be $\mathcal{M}$. $\bar{\Theta}'$ is the subset of Θ which satisfies the following condition. The first index in superscripts for $\mathcal{M}$ represents the residual block index and the second represents the layer index. The first index in subscripts represents the convolutional matrix index and the second is for notational distinction. $\bar{\Theta}' \subset \Theta$ is a set of parameters defined as follows.

1. For $i = 1, \cdots, N^*$ and $\ell = 1, \cdots, L$,

$$w^{(i,\ell)}_{p,:,:} = \begin{pmatrix} w^{(i,\ell)^*}_{p,:,:} & 0 \\ \mathcal{M}^{(i,\ell)}_{p,1} & \mathcal{M}^{(i,\ell)}_{p,2} \end{pmatrix}, \quad b^{(i,\ell)} = \begin{pmatrix} b^{(i,\ell)^*} \\ \mathcal{M}^{(i,\ell)} \end{pmatrix}$$

   for $p = 1, \cdots, K$.

2. For $i = N^* + 1, \cdots, N$ and $\ell = 1, \cdots, L - 1$,

$$w^{(i,\ell)}_{p,:,:} = \mathcal{M}^{(i,\ell)}_{p,1}, \quad b^{(i,\ell)} = \mathcal{M}^{(i,\ell)}$$

   for $p = 1, \cdots, K$, where $\mathcal{M}^{(i,\ell)}_{p,1} \in \mathbb{R}^{M \times M}$ and $\mathcal{M}^{(i,\ell)} \in \mathbb{R}^{M}$.

3. For $i = N^* + 1, \cdots, N$ and $\ell = L$,

$$w^{(i,L)}_{p,:,:} = \begin{pmatrix} 0_{M^* \times M} \\ \mathcal{M}^{(i,L)}_{p,1} \end{pmatrix}, \quad b^{(i,L)} = \begin{pmatrix} 0_{M^*} \\ \mathcal{M}^{(i,L)} \end{pmatrix}$$

   for $p = 1, \cdots, K$, where $0_{M^* \times M} \in \mathbb{R}^{M^* \times M}$ and $0_{M^*} \in \mathbb{R}^{M^*}$ denote a zero matrix and zero vector respectively.

4. For the fully-connected layer,

$$w = \begin{pmatrix} w^* & 0 \end{pmatrix}, \quad b = b^*.$$

Now we show that $\bar{\Theta}'$ is a subset of the teacher equivalent set.

Lemma 14. *$\bar{\Theta}'$ is a subset of $\bar{\Theta}$.*

Proof. We show that the student model with parameters in $\bar{\Theta}'$ exactly replicates the teacher model.

Let the output from the ℓ-th convolutional layer of the i-th residual block in the teacher model and the student model be $h^{(i,\ell)^*} \in \mathbb{R}^{m \times M^*}$ and $h^{(i,\ell)} \in \mathbb{R}^{m \times M}$ respectively. In addition, let the output from the i-th residual block after adding an identity mapping in the teacher model and student model be $h^{(i)^*} \in \mathbb{R}^{m \times M^*}$ and $h^{(i)} \in \mathbb{R}^{m \times M}$ respectively.

First, we show that $h^{(N^*)}_{:,1:M^*} = h^{(N^*)^*}$. For $i = 1$, $\ell = 1$, we have for $k = 1, \cdots, m$

$$\begin{aligned} h^{(1,1)}_{k,:} =& \sigma \left(\sum_{p=1}^{K} w^{(1,1)}_{p,:,:} \begin{pmatrix} x_{k+p-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + b^{(1,1)} \right) \\ =& \begin{pmatrix} h^{(1,1)^*}_{k,:} \\ g^{(1,1)}_{k} \end{pmatrix}, \end{aligned}$$

where $g_k^{(1,1)} \in \mathbb{R}^{M-M^*}$ denotes elements in the redundant width. For $i = 1$ and $\ell = 2$, we have for $k = 1, \cdots, m$

$$\begin{aligned} h_{k,:}^{(1,2)} =& \sigma\left(\sum_{p=1}^{K} w_{p,:,:}^{(1,2)} h_{k+p-1,:}^{(1,1)} + b^{(1,2)}\right) \\ =& \begin{pmatrix} h_{k,:}^{(1,2)^*} \\ g_k^{(1,2)} \end{pmatrix}, \end{aligned}$$

where $g_k^{(1,2)} \in \mathbb{R}^{M-M^*}$ denotes elements in the redundant width. Continuing this recursively, we have for $k = 1, \cdots, m$

$$h_{k,:}^{(1,L)} = \begin{pmatrix} h_{k,:}^{(1,L)^*} \\ g_k^{(1,L)} \end{pmatrix},$$

where $g_k^{(1,L)} \in \mathbb{R}^{M-M^*}$ denotes elements in the redundant width. After adding the identity mapping, we have for $k = 1, \cdots, m$

$$\begin{aligned} h_{k,:}^{(1)} =& \begin{pmatrix} x_k \\ 0_{M-1} \end{pmatrix} + \begin{pmatrix} h_{k,:}^{(1,L)^*} \\ g_k^{(1,L)} \end{pmatrix} \\ =& \begin{pmatrix} h_{k,:}^{(1)^*} \\ g_k^{(1)} \end{pmatrix}, \end{aligned}$$

where $0_{M-1} \in \mathbb{R}^{M-1}$ denotes a zero vector and $g_k^{(1)} \in \mathbb{R}^{M-M^*}$ denotes elements in the redundant width. Continuing this argument recursively for $i = 1, \cdots, N^*$, we have

$$h_{k,:}^{(N^*)} = \begin{pmatrix} h_{k,:}^{(N^*)^*} \\ g_k^{(N^*)} \end{pmatrix},$$

where $g_k^{(N^*)} \in \mathbb{R}^{M-M^*}$ denotes elements in the redundant width. Hence, we obtain $h_{:,1:M^*}^{(N^*)} = h^{(N^*)^*}$.

Second, we show that $h_{:,1:M^*}^{(N)} = h^{(N^*)^*}$. For $i = N^* + 1$ and $\ell = L$, we have for $k = 1, \cdots, m$

$$\begin{aligned} h_{k,:}^{(N^*+1,L)} =& \sigma\left(\sum_{p=1}^{K} w_{p,:,:}^{(N^*+1,L)} h_{k+p-1,:}^{(N^*+1,L-1)} + b^{(N^*+1,L)}\right) \\ =& \begin{pmatrix} 0_{M^*} \\ g_k^{(N^*+1,L)} \end{pmatrix} \end{aligned}$$

since $\sigma(0) = 0$, where $0_{M^*} \in \mathbb{R}^{M^*}$ denotes a zero vector and $g_k^{(N^*+1,L)} \in \mathbb{R}^{M-M^*}$ denotes elements in the redundant width. Hence, after adding the identity mapping from the previous residual block, we have for $k = 1, \cdots, m$

$$\begin{aligned} h_{k,:}^{(N^*+1)} =& \begin{pmatrix} h_{k,:}^{(N^*)^*} \\ g_k^{(N^*)} \end{pmatrix} + \begin{pmatrix} 0_{M^*} \\ g_k^{(N^*+1,L)} \end{pmatrix} \\ =& \begin{pmatrix} h_{k,:}^{(N^*)^*} \\ g_k^{(N^*+1)} \end{pmatrix}, \end{aligned}$$

where $g_k^{(N^*+1)} \in \mathbb{R}^{M-M^*}$ denotes elements in the redundant width. Continuing this recursively for $i = N^* + 1, \cdots, N$, we finally obtain $h_{:,1:M^*}^{(N)} = h^{(N^*)^*}$.

Finally, we show that the final output from the fully connected layer in the student model is the same as that in the teacher model. The output from the student model after the flatten operation is

$$\begin{pmatrix} h_{1,1}^{(N)} \\ \vdots \\ h_{m,1}^{(N)} \\ \vdots \\ h_{1,M}^{(N)} \\ \vdots \\ h_{m,M}^{(N)} \end{pmatrix} = \begin{pmatrix} {h_{1,1}^{(N^*)}}^* \\ \vdots \\ {h_{m,1}^{(N^*)}}^* \\ \vdots \\ {h_{1,M^*}^{(N^*)}}^* \\ \vdots \\ {h_{m,M^*}^{(N^*)}}^* \\ g_k^{(N)} \end{pmatrix},$$

where $g_k^{(N)} \in \mathbb{R}^{m(M-M^*)}$ denotes elements in the redundant width. Hence the final output from the fully connected layer is

$$\begin{pmatrix} w^* & 0 \end{pmatrix} \begin{pmatrix} {h_{1,1}^{(N^*)}}^* \\ \vdots \\ {h_{m,1}^{(N^*)}}^* \\ \vdots \\ {h_{1,M^*}^{(N^*)}}^* \\ \vdots \\ {h_{m,M^*}^{(N^*)}}^* \\ g_k^{(N)} \end{pmatrix} + b^* = w^* \begin{pmatrix} {h_{1,1}^{(N^*)}}^* \\ \vdots \\ {h_{m,1}^{(N^*)}}^* \\ \vdots \\ {h_{1,M^*}^{(N^*)}}^* \\ \vdots \\ {h_{m,M^*}^{(N^*)}}^* \end{pmatrix} + b^*,$$

which is the same as the final output from the teacher model. □

We calculate the lower bound for the dimension of $\bar{\Theta}$ by counting the dimension of $\bar{\Theta}'$.

Lemma 15. *We have*

$$d_{\bar{\Theta}} \geq d_{\Theta} - d',$$

where

$$d' = N^*LM^*(KM+1) + (N-N^*)M^*(KM+1) + (mM+1)p.$$

Proof. From Lemma 14, $\bar{\Theta}$ contains $\bar{\Theta}'$. The number of the restricted parameters in $\bar{\Theta}'$ is

$$\begin{aligned} d' :=& N^*LKM^*M + N^*LM^* + (N-N^*)KM^*M + (N-N^*)M^* + mMp + p \\ =& N^*LM^*(KM+1) + (N-N^*)M^*(KM+1) + (mM+1)p. \end{aligned}$$

Hence, $\bar{\Theta}$ contains a $(d_{\Theta} - d')$-dimensional hypercube. A $(d_{\Theta} - d')$-dimensional hypercube contains an analytic manifold of dimension $d_{\Theta} - d'$. Hence the dimension of $\bar{\Theta}$ is at least $d_{\Theta} - d'$, which completes the proof. □

Substituting the result of Lemma 15 into Theorem 2, we obtain $\widehat{k} \leq d' + 1$, which completes the proof of Theorem 5.

G Details of Experiments

G.1 Guess and Check algorithm for Sampling Random Near Interpolators

To obtain a predictor with a random near interpolator, we can utilize the Guess & Check (G&C) algorithm by Chiang et al. (2023), which provides a practical implementation to sample random near interpolators. Specifically, the G&C algorithm is operated by the following procedure:

1. Fix a sufficiently small $\varepsilon > 0$.
2. Sample a parameter θ according to the uniform distribution on Θ, without considering the training set.
3. For the sampled parameter θ, evaluate the training loss $L_n(\theta)$ using the training set $\{(x_i, y_i)\}_{i=1}^n$.
4. If $L_n(\theta) \leq \varepsilon$ holds, return θ; otherwise, go back to 2.

In our experiment in Section 6, we set $\varepsilon = 0.01$. Moreover, we sample a parameter θ according to Xavier uniform distribution as step 2.

H Additional Experiments

H.1 Near Interpolators in Large Models on MNIST

In this section, we present a complementary experiment on MNIST using LeNet in order to illustrate the behavior in a more practical classification setting.

Setup. We study the generalization behavior of near interpolators for LeNet (LeCun et al., 1998) on the MNIST dataset.

To generate near interpolators, we train LeNet using Xavier initialization (Glorot & Bengio, 2010), a cross-entropy loss, and the Adam optimizer with batch size 1024 and learning rate 0.001. Training is continued until the full-batch training loss on the selected MNIST subset falls below 0.01. This procedure is repeated 2000 times, yielding 2000 near interpolator samples.

To estimate the dimension $d_{\bar{\Theta}}$ of the TES, we repeatedly train LeNet on the full MNIST dataset using Adam with batch size 1024 until the full-batch training loss falls below 0.01. This process is repeated 30000 times, yielding 30000 approximate teacher-equivalent parameter samples.

We then estimate the dimension of $\bar{\Theta}$ by utilizing the lPCA algorithm (Fukunaga & Olsen, 1971), implemented by the scikit-dimension package (Bac et al., 2021).

We compute the test loss on the MNIST test split. In order to reproduce the teacher-student setting on the MNIST dataset, we remove about 9 % of ambiguous data from the test split of MNIST by utilizing the cleanlab package (Northcutt et al., 2020).

Remark 3 (Algorithm-dependent bias)**.** As discussed in Section 6, our theory is intended to describe distributions over interpolators that are not tied to a specific optimization algorithm. Ideally, one would therefore employ a sampling procedure such as the Guess and Check algorithm. However, this is computationally infeasible for large-scale models, and we instead use Adam as a practical approximation.

To evaluate the robustness of the observed behavior, we additionally compare against the *pattern search* algorithm, which serves as an alternative sampling procedure designed to reduce optimizer-dependent bias. The results are qualitatively consistent across these algorithms. Details are provided in Appendix H.2.

Remark 4 (Non-analytic activation functions)**.** Our theoretical analysis assumes analyticity of the activation function, whereas the experiments use ReLU activations in LeNet. Despite this mismatch, we observe qualitatively similar behavior, suggesting that the geometric phenomenon studied in this paper may extend beyond analytic settings.

Result. Figure 3 shows the test losses of random near interpolators for LeNet on MNIST. The estimated upper bound on the strong sample complexity, indicated by the red vertical line, is qualitatively consistent with the observed transition in generalization behavior. In particular, the bound approximately captures the sample scale at which near interpolators begin to achieve low test loss.

As expected from Proposition 3, the test losses do not converge exactly to zero because the sampled parameters are near interpolators rather than exact interpolators.

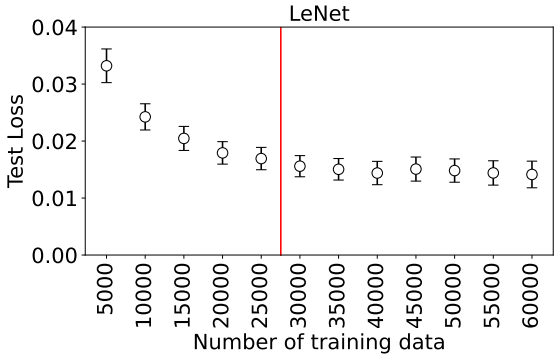

Figure 3: Test losses of random near interpolators on LeNet. The vertical axis represents the test loss, while the horizontal axis corresponds to the number of training data. The error bars indicate the standard deviation over 2000 trials for each training sample size. The red vertical line is the estimated upper bound of the strong sample complexity $d_\Theta - d_{\bar{\Theta}} + 1$.

H.2 Experiment by pattern search

For an additional robustness check against optimizer-specific effects, we sample random interpolators using the pattern search algorithm and examine their properties. The method is introduced in Chiang et al. (2023) as an alternative to the G&C algorithm that is free from the implicit bias associated with gradient-based methods. Specifically, the algorithm perturbs the parameter in a random direction and then evaluates whether this perturbation decreases the objective function. If the update results in a decrease, the new parameter is accepted; otherwise, the parameter is reverted to its original value. Furthermore, if none of the attempted updates yield a decrease, we reduce the step size accordingly. A more detailed description of the procedure is provided in Algorithm 1.

Algorithm 1: Pattern Search Algorithm

Input: Initial point θ_0, step size $\alpha_0 > 0$, stopping threshold ε
Output: Approximate solution θ_k
Set $k \leftarrow 0$;
while $L_n(\theta_k) > \varepsilon$ **do**
  Compute trial points $\theta_k + \alpha_k e_{i(k)}$ or $\theta_k - \alpha_k e_{i(k)}$, where $i(k) \in \{1, \cdots, d_\Theta\}$ is randomly chosen and $e_{i(k)}$ denotes a vector whose $i(k)$-th component is 1 and all other components are 0;
  if *there exists* $i(k) \in \{1, \cdots, d_\Theta\}$ *s.t.* $L_n(\theta_k + \alpha_k e_{i(k)}) < L_n(\theta_k)$ **then**
    $\theta_{k+1} \leftarrow \theta_k + \alpha_k e_{i(k)}$;
    $\alpha_{k+1} \leftarrow \alpha_k$;
  else if *there exists* $i(k) \in \{1, \cdots, d_\Theta\}$ *s.t.* $L_n(\theta_k - \alpha_k e_{i(k)}) < L_n(\theta_k)$ **then**
    $\theta_{k+1} \leftarrow \theta_k - \alpha_k e_{i(k)}$;
    $\alpha_{k+1} \leftarrow \alpha_k$;
  else
    $\theta_{k+1} \leftarrow \theta_k$;
    $\alpha_{k+1} \leftarrow \gamma_{\mathrm{dec}}\, \alpha_k$;
  $k \leftarrow k + 1$;

We describe the details of our experimental setup. Because this algorithm is computationally expensive, we restrict the MNIST dataset to samples whose labels are 0 or 1. We employ a 2-layer FCDNN with a hidden-layer width of 10 and the softplus activation function. The loss function is the cross-entropy loss. The model parameters are initialized using Xavier uniform initialization, and the step size is initialized as 1.0. We set the stopping threshold to $\varepsilon = 0.01$.

To estimate the dimension $d_{\bar{\Theta}}$ of $\bar{\Theta}$, we approximately sample from $\bar{\Theta}$ by training the 2-layer FCDNN on all MNIST samples labeled 0 or 1 using the pattern search algorithm. This procedure is repeated 10000 times, producing 10000 approximate samples from $\bar{\Theta}$. We then estimate the dimension of the manifold on which these samples lie by applying the lPCA algorithm implemented in the scikit-dimension package (Bac et al., 2021).

Next, we generate 1000 random near interpolators using the pattern search algorithm, varying the number of training samples across $1000, 2000, 3000, 4000, 5000, 6000$, and 7000, and compute their test loss on the MNIST test split.

The results are presented in Figure 4. The estimated upper bound is consistent with the empirical findings, as it provides a sufficient number of samples for the generalization of random near interpolators.

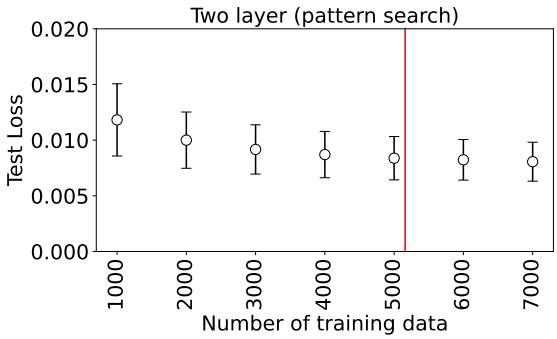

Figure 4: Test losses of random near interpolators on 2-layer FCDNN. The vertical axis represents the test loss, while the horizontal axis corresponds to the number of training data. The error bars indicate the standard deviation over 1000 trials for each training sample size. The red vertical line is the estimated upper bound of the strong sample complexity $d_{\Theta} - d_{\bar{\Theta}} + 1$.