How Many Data Are Needed for Robust Learning?

Hongyang Zhang*  Yihan Wu†  Heng Huang‡

February 24, 2022

Abstract

We show that the sample complexity of robust interpolation problem could be exponential in the input dimensionality and discover a phase transition phenomenon when the data are in a unit ball. Robust interpolation refers to the problem of interpolating \( n \) noisy training data in \( \mathbb{R}^d \) by a Lipschitz function. Although this problem has been well understood when the covariates are drawn from an isoperimetry distribution, much remains unknown concerning its performance under generic or even the worst-case distributions. Our results are two-fold: 1) too many data hurt robustness; we provide a tight and universal Lipschitzness lower bound \( \Omega(n^{1/d}) \) of the interpolating function for arbitrary data distributions. Our result disproves potential existence of an \( O(1) \)-Lipschitz function in the overparametrization scenario when \( n = \exp(\omega(d)) \). 2) Small data hurt robustness: \( n = \exp(\Omega(d)) \) is necessary for obtaining a good population error under certain distributions by any \( O(1) \)-Lipschitz learning algorithm. Perhaps surprisingly, our results shed light on the curse of big data and the blessing of dimensionality for robustness, and discover an intriguing phenomenon of phase transition at \( n = \exp(\Theta(d)) \).

1 Introduction

Robustness has been a central research topic in machine learning [SZS+14, GSS14], statistics [Hub04], operation research [BTEGN09], and many other domains. In machine learning, study of adversarial robustness has led to significant advances in defending against adversarial attacks, where test inputs with slight modification can lead to problematic prediction results. In statistics and operation research, robustness is a desirable property for optimization problems against uncertainty, which can be represented as deterministic or random variability in the value of optimization parameters. This is known as robust statistics or robust optimization. In both cases, the problem can be stated as given a deterministic labeling function \( g : \mathbb{R}^d \to [-1, 1] \), (approximately) interpolating the training data \( \{(x_i, g(x_i))\}_{i=1}^n \) or its noisy counterpart by a function with small Lipschitz constant. The focus of this paper is on the latter setting known as robust interpolation problem [BS21]. That is, given noisy training data \( \{(x_i, g(x_i) + z_i)\}_{i=1}^n \) of size \( n \) where \( x_1, \ldots, x_n \) are restricted in a unit ball and \( z_1, \ldots, z_n \) have variance \( > 0 \), how many data are needed for robust learning provided that the functions in the class can (approximately) interpolate the noisy training data with Lipschitz constant \( L \)?

There are several reasons to study the noisy setting [BS21]: 1) The real-world data are noisy. For example, it has been shown that around 3.3% of the data in the most-cited datasets was inaccurate or mislabeled [NAM21]. 2) This noise assumption is necessary from a theoretical point of view, as otherwise there could exist a Lipschitz function which perfectly fits the training data for any large \( n \). Despite progress on the robust interpolation problem [BS21, BLN21], many fundamental questions remain unresolved. In modern learning theory, it was commonly believed that 1) big data [SST+18], 2) low dimensionality of input [BDMZ20, YDH+20, KLGF20],
and 3) overparametrization [BS21, BLN21] improve robustness. We view the robustness problem from the perspective of Lipschitzness and ask the following question:

Are big data and large models a remedy for robustness?

In fact, there is significant evidence to indicate that enlarging the data amount and model size improves robustness when \( n \) is moderately large (e.g., when \( n = \text{poly}(d) \), see [MMS+17, SST+18]). Our work goes further by showing that for any learning algorithm, there exists a joint data distribution such that one needs at least \( \exp(\Omega(d)) \) samples to learn an \( \mathcal{O}(1) \)-Lipschitz function. On the other hand, we show that big data and large models may not be a remedy for robustness if \( n \) goes even larger, in particular, \( n \geq \exp(\omega(d)) \). Our main contribution is summarized by a sketch in Figure 1, which discovers \( \exp(\Theta(d)) \) as the sample complexity when \( L = \mathcal{O}(1) \) and reveals a phase transition phenomenon of astuteness\(^1\) with different amounts of training data.

The robust interpolation problem becomes more challenging when no assumptions are made on the distribution of covariates. Due to the well-separated nature of data, most positive results for obtaining good Lipschitzness lower bound have focused on the isoperimetry distribution [BS21, BLN21]. Isoperimetry states that the output of any Lipschitz function is \( \mathcal{O}(1) \)-subgaussian under suitable rescaling. Special cases of isoperimetry include high-dimensional Gaussians \( \mathcal{N}(0, I_d) \), uniform distributions on spheres and hypercubes of diameter 1. However, real-world data might not follow the isoperimetry assumption. In contrast, our results of Theorem 3.1 go beyond isoperimetry and provide a universal lower bound of robustness for any model class, including the class of neural networks with arbitrary architecture, under arbitrary distributions in the bounded space.

### 1.1 Our results

Our results consist of two parts: a) too many data hurt robustness (Section 3), and b) small data hurt robustness (Section 4). The two results are complementary and reveal \( \exp(\Theta(d)) \) as the sample complexity for \( \mathcal{O}(1) \)-Lipschitz learning in the robust interpolation problem.

**Too many data hurt robustness.** Lipschitzness (or local Lipschitzness) is an important characterization of adversarial robustness for learning algorithms [YRZ+20, ZYJ+19]. For a given score function \( f \), we denote by \( \text{Lip}_{\|\cdot\|}(f) \) the Lipschitz constant of \( f \) w.r.t. the norm \( \|\cdot\| \). That is, for any \( x_1, x_2 \) in the input space,\(^1\)
Theorem 3.2 (informal, Lipschitzness lower bound). Let $\mathcal{F}$ be any class of functions from $\mathbb{R}^d \to [-1, 1]$ and let $\{(x_i, y_i)\}_{i=1}^n$ be i.i.d. input-output pairs in $\{x : \|x\| \leq 1\} \times [-1, 1]$ for any given norm $\|\cdot\|$. Assume that the expected conditional variance of the output (i.e., the “noise level”) is strictly positive, denoted by $\sigma^2 := \mathbb{E} \text{Var}[y|x] > 0$. Then, with high probability over the sampling of the data, one has simultaneously for all $f \in \mathcal{F}$:
\[
\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \Rightarrow \text{Lip}_{\|\cdot\|}(f) \geq \Omega(\epsilon n^{1/d}).
\]

Intuitively, the Lipschitzness of the interpolating function is inversely propositional to the distance between the closest training data pairs. Given $n$ training data in the $d$-dimensional bounded space, one can scatter the data by the training distribution in Figure 2, where the distance between any training pair is as large as $\Theta(1/n^{1/d})$. Inspired by this, we complement Theorem 3.1 with a matching Lipschitzness upper bound of $O(n^{1/d})$, which shows that the Lipschitz lower bound in Theorem 3.1 is achievable by a certain function:

Theorem 3.1 (informal, Lipschitzness lower bound). Let $\mathcal{F}$ be any class of functions from $\mathbb{R}^d \to [-1, 1]$ and let $\{(x_i, y_i)\}_{i=1}^n$ be i.i.d. input-output pairs in $\{x : \|x\| \leq 1\} \times [-1, 1]$ for any given norm $\|\cdot\|$. Assume that the expected conditional variance of the output (i.e., the “noise level”) is strictly positive, denoted by $\sigma^2 := \mathbb{E} \text{Var}[y|x] > 0$. Then, with high probability over the sampling of the data, one has simultaneously for all $f \in \mathcal{F}$:
\[
\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \Rightarrow \text{Lip}_{\|\cdot\|}(f) \geq \Omega(\epsilon n^{1/d}).
\]

Figure 2: The distribution of training data that achieves the Lipschitzness lower bound of Theorem 3.1. Without loss of generality, let $\sigma^2$ and $\epsilon$ be absolute constants. The existence of noise implies $|y_i - y_j| = \Theta(1)$ for any $i, j$ with high probability. Since the distance between any two nearby points is as small as $O(1/n^{1/d})$, the Lipschitz constant of $f$ which perfectly fits the noisy targets would be $\Omega(n^{1/d})$. 
\[
|f(x_1) - f(x_2)| \leq \text{Lip}_{\|\cdot\|}(f)\|x_1 - x_2\|. 
\]

Our first result derives a (tight) lower bound on the Lipschitzness of learned functions when the training error is slightly smaller than the noise level (i.e., in the case of overfitting), but without assumptions on the distribution of covariates except that they are restricted in the bounded space $\mathcal{X} := \{x : \|x\| \leq 1\}$. We are interested in the assumption of bounded space because: 1) most applications of machine learning focus on the case where the data are in the bounded space. For example, images and videos are considered to be in $[-1, 1]^d$. 2) The discussion of Lipschitzness is closely related to how large the input space is. For example, for the images restricted in $[-1, 1]^d$, special attentions are paid on the $\ell_\infty$ robust radius of 0.031 or 0.062 [ZYJ19, MMS17], which corresponds to a (local) Lipschitz constant of $O(1)$ for the classifier.

The universal law of robustness by [BS21] provides an $\tilde{\Omega}(\sqrt{nd/p})$ Lipschitzness lower bound of the interpolating functions when the underlying distribution is isoperimetry (see Theorem 2.1). While their results predict potential existence of an $O(1)$-Lipschitz function satisfying the “overfitting” condition when $p = \Omega(nd)$, we disprove the hypothesis in the big data scenario when $n = \exp(\omega(d))$ for arbitrary distributions. Our main results are as follows (the detailed theorems are introduced at later sections):

Theorem 3.1 (informal, Lipschitzness lower bound). Let $\mathcal{F}$ be any class of functions from $\mathbb{R}^d \to [-1, 1]$ and let $\{(x_i, y_i)\}_{i=1}^n$ be i.i.d. input-output pairs in $\{x : \|x\| \leq 1\} \times [-1, 1]$ for any given norm $\|\cdot\|$. Assume that the expected conditional variance of the output (i.e., the “noise level”) is strictly positive, denoted by $\sigma^2 := \mathbb{E} \text{Var}[y|x] > 0$. Then, with high probability over the sampling of the data, one has simultaneously for all $f \in \mathcal{F}$:
\[
\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \Rightarrow \text{Lip}_{\|\cdot\|}(f) \geq \Omega(\epsilon n^{1/d}).
\]
\( f^*(x_i) = y_i, \forall i \in [n], \) then use the linear interpolation between neighbour training points as the prediction of other samples. This function is at most \( 2n^{1/d} \)-Lipschitz.

**Small data hurt robustness.** The lower bound given by Theorem 3.1 implies that one can sample at most \( \exp(O(d)) \) training data in order to obtain an \( O(1) \)-Lipschitz function in the robust interpolation problem. The following result shows that \( n = \exp(O(d)) \) is a necessary condition for obtaining a good population error by any \( O(1) \)-Lipschitz learning algorithm:

**Theorem 4.1** (informal, sample complexity lower bound). Let \( S = \{(x_i, y_i)\}_{i=1}^n \) be i.i.d. training pairs in \( \{x : \|x\| \leq 1\} \times [-1, 1] \) for any given norm \( \|\cdot\| \). Denote by \( \mathcal{L}_D(f) := \mathbb{E}_D[(f(x) - y)^2] \) the squared \( \ell_2 \) loss. Assume that the expected conditional variance of the output (i.e., the “noise level”) is strictly positive and bounded by \( 1/2 \), denoted by \( \sigma^2 := \mathbb{E}[\text{Var}[y|x]] \). Let \( A(S) : \mathcal{X} \rightarrow \mathbb{R} \) be any \( L \)-Lipschitz learning algorithm over a training set \( S \). Then there exists a distribution \( D' \) of \( (x, y) \) such that

\[
\begin{align*}
n < \frac{1}{2} \left( \frac{2L}{1-2\epsilon} \right)^d \implies \mathbb{E}_S[\mathcal{L}_{D'}(A(S))] \geq \min \left\{ \frac{1}{4}, \epsilon \right\} + \sigma^2.
\end{align*}
\]

As a consequence of Theorems 3.1 and 4.1, we obtain the sketch of phase transition in Figure 1. The results in this paper hold true for any function class \( \mathcal{F} \), including the class of neural networks with arbitrary architecture.

### 1.2 Our (counter-intuitive) implications

It was widely believed that 1) big data [SST+18], 2) low dimensionality of input [BDMZ20], and 3) overparametrization [BS21, BLN21, GCL+19] improve robustness. Our main results of Theorem 3.1 challenge the common beliefs and show that these hypotheses may not be true in the robust interpolation problem. Our results shed light on the theoretic understanding of robustness beyond isoperimetry assumption.

**The curse of big data.** Our Lipschitzness lower bound in Theorem 3.1 is increasing w.r.t. the sample size \( n \). The intuition is that as one has more training data, those data are squeezed in the bounded space with smaller margin. Thus to fit the data well, the Lipschitz constant of the interpolating functions cannot be small. Perhaps surprisingly, our results contradict with the common belief that more data always improve model robustness.

**The blessing of dimensionality.** It is known that high dimensionality of input space strengthens the power of adversary. For example, in the \( \ell_{\infty} \) threat model, an adversary can change every pixel of a given image by 8 or 16 intensity levels. Admittedly, higher dimensionality means that the adversary can modify more pixels. However, we show that our Lipschitzness lower bound in Theorem 3.1 is decreasing w.r.t. \( d \). The intuition is that input space with higher dimension has larger space to scatter the data. So the data can be well-separated, and thus the Lipschitz constant of the interpolating functions can be small.

**Overfitting (due to overparametrization) may hurt robustness.** Our Lipschitzness lower bound in Theorem 3.1 is increasing w.r.t. the overfitting level \( \epsilon \). In fact, one can show that \( \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \) implies \( p = \tilde{\Omega}(n\epsilon^2) \) for all \( f \in \mathcal{F} \) with high probability, if \( \mathcal{F} \) admits a Lipschitz parametrization by \( p \) real parameters (see Theorem D.4). Thus, larger \( p \) allows larger \( \epsilon \), which leads to larger Lipschitz constant according to Theorem 3.1. The argument supports a contemporaneous independent work of [HJ22] on the curse of overparametrization in adversarial training and verifies an overfitting concern in the robust deep learning [RWK20].

### 1.3 Technical overview

We discuss the techniques used for achieving our results.
Lipschitz generalization bound (Theorem 3.1). We begin our analysis with a key observation that the training error of \( f \) is smaller than the best possible population error (i.e., the noise level) with an \( \epsilon \) margin. This implies that the generalization error is at least \( \epsilon \). Thus, the remainder of the proof is to relate the generalization error to the Lipschitzness of functions by a uniform convergence bound. Our proof is based on the relation between generalization bound and Rademacher complexity of the given function space \( \mathcal{F} \) and the error function \( l \). With a large probability, the generalization error is bounded by the Rademacher complexity of \( l \circ \mathcal{F} := \{ l(f) : f \in \mathcal{F} \} \) plus an \( \mathcal{O}(1/\sqrt{n}) \) term. So if we assume the generalization error is at least \( \epsilon \), the Rademacher complexity of \( l \circ \mathcal{F} \) has to be of order \( \Theta(\epsilon) \). When \( l \circ f = l(f(x), y) := (f(x) - y)^2 \), we prove that the Rademacher complexity of \( l \circ \mathcal{F} \) is upper bounded by \( 4 \) times of the Rademacher complexity of \( \mathcal{F} \). The Rademacher complexity of \( \mathcal{F} \) is directly related to the covering number of \( \mathcal{F} \). Inspired by [KT61], we show that if \( \forall f \in \mathcal{F}, \text{Lip}_{\| \cdot \|}(f) \leq L \), the \( \eta \)-covering number of \( \mathcal{F} \) is bounded by \( \left\lceil \frac{2L \text{diam}(\mathcal{X})}{\eta} \right\rceil 2^N \), where \( \mathcal{X} \) is the input space and \( N \) is the number of \( \frac{n}{2\pi} \)-covering of \( \mathcal{X} \). We show that if \( L = o(n^{1/d}) \), the Rademacher complexity of \( \mathcal{F} \) is \( o(\epsilon) \), which contradicts to our assumption that the generalization error is larger than \( \epsilon \). Thus all functions in \( \mathcal{F} \) must have Lipschitz constants of order \( \Omega(n^{1/d}) \), which yields Theorem 3.1.

No-free-lunch theory (Theorem 4.1). We aim to find a bad distribution \( \mathcal{D}' \) such that for all \( \mathcal{O}(1) \)-learning algorithms with a set of training samples \( S \) of size \( \exp(\mathcal{O}(d)) \), the expected error on \( \mathcal{D}' \) over the sampling of \( S \) is strictly larger than the best possible error \( \sigma^2 \). The main idea to find such a distribution is that any learning algorithms, which can only have access to a half of the training set, has no information about the rest part. Consider the set of distributions \( \{ \mathcal{D}_i \} \) on a sample set \( C \), where \( |C| = \exp(\mathcal{O}(d)) \), whose targets on the “rest part” are contradictory, e.g., \( (x, y = 1) \in \mathcal{D}_0 \), \( (x, y = -1) \in \mathcal{D}_1 \). That is, \( \mathcal{D}_0 \) and \( \mathcal{D}_1 \) have contradicted targets on \( x \). Thus the averaged error of an arbitrary learning algorithm \( \mathcal{A}(S) \) over the set of distributions \( \{ \mathcal{D}_i \} \) is non-zero and we denote it by \( \epsilon \). We then consider the \( \mathcal{D}_{\epsilon^i} \in \{ \mathcal{D}_i \} \) which maximizes the error of \( \mathcal{A}(S) \). Obviously, the error of \( \mathcal{A}(S) \) on \( \mathcal{D}_{\epsilon^i} \) is larger than the averaged error \( \epsilon \) over \( \{ \mathcal{D}_i \} \), which leads to Theorem 4.1.

1.4 Notations

We will use \( \mathcal{X} \) to represent the instance space, \( \mathcal{F} = \{ f : \mathcal{X} \to [-1, 1] \} \) to represent the hypothesis/function space, \( x \in \mathcal{X} \) to represent the sample instance, \( y \in [-1, 1] \) to represent the target, and \( z \) to represent the target noise. For errors, denote by \( l(f(x), y) \) the loss function of \( f \) on instance \( x \) and target \( y \). Let \( \mathcal{L}_D(f) := \mathbb{E}_{(x,y) \sim D}[l(f(x), y)] \) be the population error, and let \( \mathcal{L}_S(f) := \frac{1}{|S|} \sum_{(x,y) \in S}[l(f(x), y)] \) be the empirical error. Denote by \( f : \mathcal{X} \to [-1, 1] \) the prediction function which maps an instance to its predicted target. It can be parameterized, e.g., by deep neural networks. For norms, we denote by \( \| \cdot \| \) a generic norm. Examples of norms include \( \| x \|_{\infty} \), the infinity norm, and \( \| x \|_2 \) the \( \ell_2 \) norm. We will frequently use \( \langle \mathcal{X}, \| \cdot \| \rangle \) to represent the normed linear space of \( \mathcal{X} \) with norm \( \| \cdot \| \). Define \( \text{diam}(\mathcal{X}) \) as the diameter of \( \mathcal{X} \) w.r.t. the norm \( \| \cdot \| \). For a given score function \( f \), we denote by \( \text{Lip}_{\| \cdot \|}(f) \) (or sometimes \( \text{Lip}(f) \) for simplicity) the Lipschitz constant of \( f \) w.r.t. the norm \( \| \cdot \| \). Let \( \lceil \cdot \rceil \) represent the ceiling operator. We will use \( \mathcal{O}(\cdot), \Theta(\cdot), o(\cdot), \) and \( \Omega(\cdot) \) to express sample complexity and Lipschitzness, and \( \mathcal{O}(\cdot) \) and \( \Omega(\cdot) \) to ignore the \( \ln(\cdot) \) factors.

2 Related Work

Robust interpolation problem. [BLN21] provided the first guarantee on the law of robustness for two-layer neural networks which was later extended by [BS21] to a universal law of robustness for general class of functions:

**Theorem 2.1** (Theorem 1 of [BS21]). Let \( \mathcal{F} \) be a class of functions from \( \mathbb{R}^d \to [-1, 1] \) and let \( \{(x_i, y_i)\}_{i=1}^n \) be i.i.d. input-output pairs in \( \mathbb{R}^d \times [-1, 1] \). Assume that:
1. \( \mathcal{F} \) admits a \( J \)-Lipschitz parametrization by \( p \) real parameters, each of size at most \( \text{poly}(n, d) \).

2. The distribution \( \mu \) of the covariates \( x_i \) satisfies isoperimetry (or a mixture thereof).

3. The expected conditional variance of the output (i.e., the “noise level”) is strictly positive, denoted by 
\[
\sigma^2 := \mathbb{E}[\text{Var}[y|x]] > 0.
\]

Then, with high probability over the sampling of the data, one has simultaneously for all \( f \in \mathcal{F} \):
\[
\frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \Rightarrow \text{Lip}_{\|\cdot\|_2}(f) \geq \tilde{\Omega} \left( \epsilon \sqrt{\frac{nd}{p \ln J}} \right).
\]

Theorems 3.1 and 2.1 share the same setting, while the former one makes much weaker assumptions: 1) Theorem 3.1 does not make any assumption on the Lipschitzness and size of model parametrization. In fact, the Lipschitzness \( J \) in Theorem 2.1 might be exponentially large w.r.t. \( n/p \) (see our discussion in Section 5.1). 2) Theorem 3.1 does not require an isoperimetry distribution of covariates. Moreover, while Theorem 2.1 predicts potential existence of an \( O(1) \)-Lipschitz function satisfying the “overfitting” condition when \( p = \tilde{\Omega}(nd) \), Theorem 3.1 disproves the hypothesis in the big data scenario when \( n = \exp(\omega(d)) \) for arbitrary distributions in the bounded space.

**Sample complexity of robust learning.** The sample complexity of robust learning for benign distributions and certain function class has been extensively studied in the recent years. In particular, [BJC21] considered the sample complexity of robust linear classification on the separated data. [YKB19] studied the adversarially robust generalization problem through the lens of Rademacher complexity. [CBM18] extended the PAC-learning framework to account for the presence of an adversary. [MHS19] showed that any hypothesis class with finite VC dimension is robustly PAC learnable with an improper learning rule. They also showed that the requirement of being improper is necessary. [SST+18] showed an \( \Omega( \sqrt{d} ) \)-factor gap between the standard and robust sample complexity for a mixture of Gaussian distributions in \( \ell_\infty \) robustness, which was later extended to the case of \( \ell_p \) robustness with a tight bound by [BCM19, DHHR20, DWR20]. Different from the prior work, our work is the first to discover the sample complexity of robust learning for arbitrary function class and learning algorithms, and discover an intriguing phenomenon of phase transition.

### 3 Too Many Data Hurt Robustness

In this section, we present our main theoretical contributions and show an intriguing observation that too many data hurt robustness.

#### 3.1 Main results

Our analysis leads to a universal lower bound of Lipschitzness regarding the robust interpolation problem.

**Theorem 3.1 (Lipschitzness Lower Bound).** Let \( \mathcal{F} \) be any class of functions from \( \mathbb{R}^d \rightarrow [-1, 1] \) and let \( \{(x_i, y_i)\}_{i=1}^{n} \) be i.i.d. input-output pairs in \( \{x : \|x\| \leq 1\} \times [-1, 1] \) for any given norm \( \|\cdot\| \). Assume that the expected conditional variance of the output (i.e., the “noise level”) is strictly positive, denoted by 
\[
\sigma^2 := \mathbb{E}[\text{Var}[y|x]] > 0.
\]

Then, with high probability over the sampling of the data, one has simultaneously for all \( f \in \mathcal{F} \):
\[
\frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \Rightarrow \text{Lip}_{\|\cdot\|_2}(f) \geq \frac{n^{1/d}}{K} \left( \frac{1}{8} \epsilon - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{n}} \right),
\]

where \( K = 96 + \frac{96 \sqrt{2 \ln n}}{d - 2} + \frac{16 \sqrt{2}}{n^{1/2} - 1/d} \sqrt{\ln(\frac{1}{3} n^{1/d} + 1)} \).
Theorem 3.1 states that, for all data distribution $\mathcal{D}$ with label noise of variance $\sigma^2$ and every function $f : \mathcal{X} \to [-1, 1]$, overfitting i.e. $\frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon$ implies $\text{Lip}_{||\cdot||} \geq \Omega(\epsilon n^{1/d})$, which indicates that achieving good astuteness is impossible when $n = \exp(\omega(d))$. In the next theorem, we show the tightness of our Lipschitzness bound.

**Theorem 3.2** (Lipschitzness Upper Bound). For any distribution $\mathcal{D}$ which is supported on $\{x \in \mathbb{R}^d : ||x|| \leq 1\}$, there exist $n$ training samples $\{x_1, ..., x_n\}$ such that $\forall i, j, i \neq j, ||x_i - x_j|| \geq \frac{1}{n^{1/d}}$. Denote by $\{y_1, ..., y_n\}$ the observed targets. We design a function $f^*$ which first perfectly fits the training samples, i.e., $f^*(x_i) = y_i, \forall i \in [n]$, then use the linear interpolation between neighbour training points as the prediction of other samples. This function is at most $2n^{1/d}$-Lipschitz.

Theorem 3.2 shows that there exist $n$ samples, such that the function which perfectly fits the training samples is $O(n^{1/d})$-Lipschitz.

### 3.2 Proof sketch of Theorem 3.1

Our proof is based on the relation between Rademacher complexity and the generalization error between the population error $\mathcal{L}_D(f)$ and the training error $\mathcal{L}_S(f)$. We defer the complete proof to Appendix A.

We begin with the definition of Rademacher complexity, which measures the richness of a function class. For a set $\mathcal{A} := \{a_1, ..., a_n\}$, the Rademacher complexity is defined as

$$R(\mathcal{A}) := \frac{1}{n} \mathbb{E}_{\sigma_1, ..., \sigma_n \in \{-1, 1\}} \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i a_i \right].$$

Given a loss function $l$, a hypothesis class $\mathcal{F}$, and a training set $S = \{(x_1, y_1), ..., (x_n, y_n)\}$, denote by $l \circ \mathcal{F} := \{l(f(\cdot)), \cdot : f \in \mathcal{F}\}$ and $l \circ \mathcal{F} \circ S := \{(l(f(x_1), y_1), ..., l(f(x_n), y_n)) : f \in \mathcal{F}\}$. The Rademacher complexity of the set $l \circ \mathcal{F} \circ S$ is given by

$$R(l \circ \mathcal{F} \circ S) := \frac{1}{n} \mathbb{E}_{\sigma_1, ..., \sigma_n \in \{-1, 1\}} \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i l(f(x_i), y_i) \right].$$

For every function $f \in \mathcal{F}$, the generation error between $\mathcal{L}_D(f)$ and $\mathcal{L}_S(f)$ is bounded by the Rademacher complexity of the function space $l \circ \mathcal{F} \circ S$. More formally, assume that $\forall f \in \mathcal{F}, \forall x \in \mathcal{X}, |l(f(x), y)| \leq a$. Then with a probability at least $1 - \delta$, for all $f \in \mathcal{F}$,

$$\mathcal{L}_D(f) - \mathcal{L}_S(f) \leq 2\mathbb{E}_{S \in \mathcal{D}^n} [R(l \circ \mathcal{F} \circ S)] + a \sqrt{\frac{2 \ln(2/\delta)}{n}}. \tag{1}$$

From Equation 1, we can see that given a lower bound of generalization gap $\mathcal{L}_D(f) - \mathcal{L}_S(f) \geq \epsilon$, one has immediately

$$\mathbb{E}_{S \in \mathcal{D}^n} [R(l \circ \mathcal{F} \circ S)] \geq \frac{\epsilon}{2} - a \sqrt{\frac{2 \ln(2/\delta)}{n}}.$$

Therefore, if we can find the relation between the Rademacher complexity of $l \circ \mathcal{F}$ and the Lipschitzness of the functions in class $\mathcal{F}$, we are able to derive a constrain of the Lipschitz constant for $\mathcal{F}$. The contraction lemma of Rademacher complexity (Lemma 26.9 of [SSBD14]) states that for a given space $A$ and a $L$-lipschitz function $h$ on $A$, we have $R(h \circ A) \leq L \cdot R(A)$. Thus, if the error function $l(f(x), y)$ is $C$-Lipschitz w.r.t. $f \in \mathcal{F}$ for arbitrary $y \in [-1, 1]$,

$$R(l \circ \mathcal{F} \circ S) \leq C \cdot R(\mathcal{F} \circ S). \tag{2}$$
It has been proved \cite{vLB04} that the Rademacher complexity of a set is directly related to the number of \( \epsilon \)-covering of the set. So the first step to calculate the Rademacher complexity of \( \mathcal{F} \circ S \) is to find the covering number of this function space.

Given a space \((\mathcal{X}, \| \cdot \|)\) and a covering radius \(\eta\), let \(N(\mathcal{X}, \eta, \| \cdot \|)\), a.k.a. the \(\eta\)-covering number, be the minimum number of \(\eta\)-ball which covers \(\mathcal{X}\). For a given function space \(\mathcal{F}\), define

\[
\|f - f'\|_\mathcal{F} = \sup_{x \in \mathcal{X}} |f(x) - f'(x)|.
\]

We have the following lemma:

**Lemma 3.3** (Covering number of \(L\)-Lipschitz function space). For a bounded and connected space \((\mathcal{X}, \| \cdot \|)\), let \(B_L\) be the set of functions \(f\)'s such that \(\text{Lip}_{\| \cdot \|}(f) \leq L\). If \(\mathcal{X}\) is connected and centered, we have for every \(\epsilon > 0\),

\[
N(B_L, \epsilon, \| \cdot \|_\mathcal{F}) \leq \left\lceil \frac{2L \cdot \text{diam}(\mathcal{X})}{\epsilon} \right\rceil 2^{N(\mathcal{X}, \frac{\epsilon}{2L}, \| \cdot \|)}.
\]

**Proof.** We consider the Lipschitz function class \(B_L := \{f \in \mathcal{F} : \text{Lip}_{\| \cdot \|}(f) \leq L\}\). In order to bound the covering number of \(\mathcal{F}\), we consider an \(\frac{\epsilon}{2L}\)-covering of input space \(\mathcal{X}\) consisting of \(N = N(\epsilon/(2L), \mathcal{X})\) plates \(U_1, U_2, ..., U_N\) centered at \(s_1, s_2, ..., s_N\). The fact that \(\mathcal{X}\) is connected enables one to join any two sets \(U_i\) and \(U_j\) by a chain of intersecting \(U_k\). For any function \(f \in \mathcal{F}\), we can construct its approximating functional \(\tilde{f}\) by taking its value on \(U_1\) as an \(\epsilon/2\)-approximation of \(f(s_1)\). As \(\text{diam}(U_1) \leq L \cdot \text{diam}(\mathcal{X})\), there are at most \(2L \cdot \text{diam}(\mathcal{X})/\epsilon\) such approximations. On the other hand, note that the \(N\) plates are chained. By Lipschitzness, the function values of \(f\) on \(s_1\) and \(s_2\) differ at most \(\epsilon/2\), and so \(\tilde{f}(s_2)\) differs at most \(\epsilon\) from \(\tilde{f}(s_1)\) by triangle inequality. It implies that to construct an \(\epsilon\)-approximation of \(f(s_2)\) on \(U_2\), we shall know either \(\tilde{f}(s_1) - \epsilon/2\) or \(\tilde{f}(s_1) + \epsilon/2\). Repeating the same argument by \(N\) times, we can bound the \(\epsilon\)-covering of \(f\) on \(\mathcal{X}\) by \(2L \cdot \text{diam}(\mathcal{X})/\epsilon 2^N\).

\cite{vLB04} proved that for every \(\epsilon > 0\),

\[
\mathbb{E}_{S' \in \mathcal{D}^n}[R(B_L \circ S)] \leq 2\epsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\epsilon/4}^{\sqrt{\text{diam}(B_L)}} \sqrt{\ln(N(B_L, u, \| \cdot \|_\mathcal{F}))} \, du.
\]

Notice that when \(u > 2L \cdot \text{diam}(\mathcal{X})\), the number of \(u\)-covering is 1 and \(\ln(N(B_L, u, \| \cdot \|_\mathcal{F})) = 0\). Combining it with Lemma 3.3 yields the following lemma:

**Lemma 3.4.** Let \((\mathcal{X}, \| \cdot \|)\) be a bounded and connected space and \(B_L\) be all functions \(f \in \mathcal{F}\) with \(\text{Lip}_{\| \cdot \|}(f) \leq L\). Let \(n = |S|\). If \(\mathcal{X}\) is connected and centered, for any \(\epsilon > 0\)

\[
\mathbb{E}_{S' \in \mathcal{D}^n}[R(B_L \circ S)] \leq 2\epsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\epsilon/4}^{2L \cdot \text{diam}(\mathcal{X})} \sqrt{N(\mathcal{X}, \frac{u}{2L}, \| \cdot \|)} \ln 2 + \ln \left[ \frac{2L \cdot \text{diam}(\mathcal{X})}{u} \right] \, du.
\]

We then have the following corollary:

**Corollary 3.5.** If \(\text{diam}(\mathcal{X}) = 2\) w.r.t. \(\| \cdot \|\) and \(d \geq 3\), we have

\[
\mathbb{E}_{S' \in \mathcal{D}^n}[R(B_L \circ S)] \leq 96 \frac{L}{n^{1/d}} + \frac{96 \sqrt{2 \ln 2}}{d - 2} \frac{L}{n^{1/d}} + 16 \sqrt{2L} \sqrt{\ln \left( \frac{1}{3} n^{1/d} + 1 \right)}.
\]

**Proof.** According to Equation 1 in \cite{MV03}, when \(\frac{u}{2L} \leq \text{diam}(\mathcal{X})\), \(N(\mathcal{X}, \frac{u}{2L}, \| \cdot \|) \leq (\frac{6L \cdot \text{diam}(\mathcal{X})}{u})^d\) if \(\mathcal{X} \subseteq \mathbb{R}^d\). Then the integral part will be \(\sqrt{(\frac{12L}{u})^d \ln 2 + \ln \left( \frac{2L \cdot \text{diam}(\mathcal{X})}{u} \right)}\), which is no more than \(\sqrt{(\frac{12L}{u})^d \ln 2 + \ln \left( \frac{4L}{u} + 1 \right)}\). Taking \(\epsilon = \Theta(\frac{L}{n^{1/d}})\), the integral part will be bounded by \(\Theta(L n^{1/2 - 1/d})\). Thus \(\mathbb{E}_{S' \in \mathcal{D}^n}[R(B_L \circ S')] \leq \Theta(\frac{L}{n^{1/d}}) + \frac{4\sqrt{2}}{\sqrt{n}} \Theta(L n^{1/2 - 1/d}) = \Theta(\frac{L}{n^{1/d}})\). We defer the detailed calculation to the appendix. \(\square\)
In our settings, we are interested in the squared $\ell_2$ loss $l(f(x), y) = (f(x) - y)^2$. We have $\nabla f(x)l(f(x), y) = 2(f(x) - y) \leq 2(|f(x)| + |y|) \leq 4$, i.e., $l(f(x), y)$ is 4-Lipschitz w.r.t. $f(x)$ for arbitrary $y \in [-1, 1]$. Thus, $\mathbb{E}_{S \in \mathcal{D}^n}[R(l \circ B_L \circ S)] \leq 4 \mathbb{E}_{S \in \mathcal{D}^n}[R(B_L \circ S)] = O\left(\frac{L}{n^{1/d}}\right)$. Combining this result with Equation 1 yields the main theorem of our paper:

**Theorem 3.6.** Let $\mathcal{X} \subseteq \mathbb{R}^d$, diam$(\mathcal{X}) = 2$, and the function space $\mathcal{F} := \{ f : \mathcal{X} \to [-1, 1]\}$. Assume that the label noise has variance $\sigma^2 = \mathbb{E}_x[\text{Var}(y|x)]$. Then with probability $1 - \delta$, for all $f \in \mathcal{F}$:

$$
\frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \Rightarrow \text{Lip}_{||\cdot||}(f) \geq \frac{n^{1/d}}{K} \left( 1 - \frac{2 \ln(2/\delta)}{n} \right),
$$

where $K = 96 + \frac{96\sqrt{2}\ln^2}{d-2} + \frac{16\sqrt{2}}{n^{2-1/d}} \sqrt{\ln(\frac{1}{3}n^{1/d} + 1)}$.

**Proof.** According to Equation 1,

$$
\mathcal{L}_D(f) - \mathcal{L}_S(f) \leq 2\mathbb{E}_{S \in \mathcal{D}^n}[R(l \circ \mathcal{F} \circ S)] + a \sqrt{\frac{2 \ln(2/\delta)}{n}},
$$

where $a := \max_{(x,y)} l(f(x), y) \leq 4$. According to Equation 2 and $\nabla f(x)l(f(x), y) \leq 4$, we have $\mathbb{E}_{S \in \mathcal{D}^n}[R([l \circ \mathcal{F} \circ S])] \leq 4 \mathbb{E}_{S \in \mathcal{D}^n}[R(\mathcal{F} \circ S)]$. Thus,

$$
\mathbb{E}_{S \in \mathcal{D}^n}[R(\mathcal{F} \circ S)] \geq \frac{1}{8} \left( \mathcal{L}_D(f) - \mathcal{L}_S(f) - 4 \sqrt{\frac{2 \ln(2/\delta)}{n}} \right).
$$

Under the label noise settings, we have

$$
\mathcal{L}_D(f) = \mathbb{E}_D[(f(x) - y)^2]
= \mathbb{E}_{x,y}[(f(x) - \mathbb{E}_y[y|x])^2 + (y - \mathbb{E}_y[y|x])^2]
\geq \mathbb{E}_x[\text{Var}(y|x)] = \sigma^2.
$$

So with the overfitting assumption $\mathcal{L}_S(f) \leq \sigma^2 - \epsilon$, we have

$$
\mathbb{E}_{S \in \mathcal{D}^n}[R(\mathcal{F} \circ S)] \geq \frac{1}{8} \left( \mathcal{L}_D(f) - \mathcal{L}_S(f) - 4 \sqrt{\frac{2 \ln(2/\delta)}{n}} \right)
= \frac{\epsilon}{8} - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{n}}.
$$

(3)

Consider $B_L = \{ f \in \mathcal{F} : \text{Lip}_{||\cdot||}(f) \leq L \}$. According to Corollary 3.5, we have

$$
K \frac{L}{n^{1/d}} \geq \mathbb{E}_{S \in \mathcal{D}^n}[R(B_L \circ S)] \geq \frac{\epsilon}{8} - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{n}},
$$

where $K = 96 + \frac{96\sqrt{2}\ln^2}{d-2} + \frac{16\sqrt{2}}{n^{2-1/d}} \sqrt{\ln(\frac{1}{3}n^{1/d} + 1)} \sim \Theta(1)$. Thus we have

$$
L \geq \frac{n^{1/d}}{K} \left( \frac{\epsilon}{8} - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{n}} \right) \sim \Omega(\epsilon n^{1/d}).
$$
If \( \exists f_0 \in \mathcal{F} \), such that
\[
L_S(f_0) \leq \sigma^2 - \epsilon \Rightarrow \text{Lip}_{\| \cdot \|}(f_0) \leq \frac{n^{1/d}}{K} \left( \frac{1}{8} \epsilon - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{n}} \right),
\]
we have
\[
\frac{\epsilon}{8} - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{n}} > K \frac{\text{Lip}_{\| \cdot \|}(f_0)}{n^{1/d}} \geq \mathbb{E}_{S \in D^n}[R(B_{\text{Lip}_{\| \cdot \|}(f_0)} \circ f) \geq \frac{\epsilon}{8} - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{n}},
\]
which yields contradiction. Therefore \( \forall f \in \mathcal{F}, \)
\[
L_S(f) \leq \sigma^2 - \epsilon \Rightarrow \text{Lip}_{\| \cdot \|}(f) \geq \frac{n^{1/d}}{K} \left( \frac{1}{8} \epsilon - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{n}} \right).
\]
Taking \( X = \{x \in \mathbb{R}^d : \|x\| \leq 1\} \), we have \( \text{diam}(X) = 2 \), which yields Theorem 3.1.

### 3.3 Proof of Theorem 3.2

**Proof.** First, we show that we can find \( n \) training samples \( \{x_1, ..., x_n\} \) such that \( \forall i, j, i \neq j, \|x_i - x_j\| \geq \frac{1}{n^{1/\sigma}} \).

Consider the \( \frac{1}{n^{1/\sigma}} \)-packing of the space \( \{x : \|x\| \leq 1\} \), the packing number is greater than the \( \frac{1}{n^{1/\sigma}} \)-covering number of the same space, which at least \( (1/n^{1/\sigma})^d = n \), we then choose \( \{x_1, ..., x_n\} \) from the \( \frac{1}{n^{1/\sigma}} \)-packing, the minimum pairwise distance is at least \( \frac{1}{n^{1/\sigma}} \). Next, we show \( f^* \) is at most \( n^{1/d} \)-Lipschitz, as \( f^* \) is the linear interpolation between neighbour training points, the worst case Lipschitz constant is \( \frac{|y_i - y_j|}{\|x_i - x_j\|} \leq 2n^{1/d} \).

### 4 Small Data Hurt Robustness

We now provide a complementary result of Section 3, which leads to a sample complexity lower bound.

#### 4.1 Main results

Our analysis leads to a sample complexity lower bound of robust interpolation problem.

**Theorem 4.1.** Let \( S = \{(x_i, y_i)\}_{i=1}^n \) be i.i.d. training pairs in \( \{x : \|x\| \leq 1\} \times [-1, 1] \) for any given norm \( \| \cdot \| \). Denote by \( L_D(f) := \mathbb{E}_D[(f(x) - y)^2] \) the squared \( \ell_2 \) loss. Assume that the expected conditional variance of the output (i.e., the “noise level”) is strictly positive and bounded by \( 1/2 \), denoted by \( \sigma^2 := \mathbb{E}[	ext{Var}[y|x]] \). Let \( A(S) : \mathcal{X} \rightarrow \mathbb{R} \) be any \( L \)-Lipschitz learning algorithm over a training set \( S \). Then there exists a distribution \( D' \) of \( (x, y) \) such that
\[
n < \frac{1}{2} \left( \frac{2L}{1 - 2\epsilon} \right)^d \Rightarrow \mathbb{E}_S[L_{D'}(A(S))] \geq \min \left\{ \frac{1}{4}, \epsilon \right\} + \sigma^2.
\]

Theorem 4.1 states that for certain distributions, the size of training samples \( n \) has to be at least \( \exp(\Omega(d)) \) if one wants to achieve good astuteness by any \( O(1) \)-Lipschitz learning algorithm. This is not restricted to the algorithms that perfectly fit the training data. The sample complexity lower bound matches the upper bound given in Theorem 3.1.
4.2 Proof sketch

The complete proof is in Appendix B. We first prove that for learning algorithms on binary classification tasks, if the number of training samples is less than half of the number of all samples, there exists a distribution with label noise such that the average error of all learning algorithms is greater than a constant. As the distribution on a binary classification is naturally a distribution on the regression tasks, we can find such a distribution for the regression tasks similarly.

Lemma 4.2. Let \( A(S) : \mathcal{X} \rightarrow \{-a, a\} \) be any learning algorithm with respect to the squared \( \ell_2 \) loss over a domain \( \mathcal{X} \) and samples \( S \). Assume there are label noise \( \mathbb{E}[\text{Var}[y|x]] = \sigma^2 \). Let \( n \) be any number smaller than \( |\mathcal{X}|/2 \), representing the size of a training set. Then, for any \( a > 0 \) there exists a distribution \( D \) (with label noise) over \( \mathcal{X} \times \{-a, a\} \) such that \( \mathbb{E}_{S \sim D^n}[\mathcal{L}_D(A(S))] \geq \frac{1}{2}(a^2 + \sigma^2) \).

In the next theorem, we will show a no-free-lunch theory on the regression tasks and algorithms that outputs an \( L \)-Lipschitz function. The intuition is to consider the minimum distance between two points in the distribution \( D \). On one hand, if the minimum distance is less than \( \epsilon \), we can assign the two samples that achieve the minimum distance with labels 1 and \(-1\), respectively. As the algorithm \( A \) is \( L \)-Lipschitz, the maximum difference between the predicted labels of the two selected points is \( L\epsilon \). Thus, the error of \( A \) will be larger than \( 1 - L\epsilon \). On the other hand, if the minimum distance is larger than \( \epsilon \), the maximum number of points in the distribution \( D \) will be less than the number of the \( \epsilon \)-packing of the input space \( \mathcal{X} \). By Lemma 4.2, there exists a distribution such that if the number of training samples is less than half of the \( \epsilon \)-packing of the input space, the average error of all learning algorithms will be at least a constant. More formally, we have the following theorem:

Theorem 4.3 (No-free-lunch theory). Let \( A(S) : \mathcal{X} \rightarrow [-1, 1] \) be any algorithm that returns an \( L \)-Lipschitz function (w.r.t. the norm \( \| \cdot \| \)) for the task of regression w.r.t. the squared \( \ell_2 \) loss over a domain \( (\mathcal{X}, \| \cdot \|) \) and samples \( S \). Let \( n \) be the size of training set, i.e., \( n = |S| \). Assume that the label noise has variance \( \sigma^2 := \mathbb{E}_D[\text{Var}(y|x)] \leq 1/2 \). Then, there exists a distribution \( D \) over \( \mathcal{X} \times [-1, 1] \) with noisy labels such that for all \( L \)-Lipschitz (w.r.t. norm \( \| \cdot \| \)) learning algorithm and any \( \epsilon \in [0, \frac{1}{2L}] \):

\[
n < M(\mathcal{X}, \epsilon, \| \cdot \|)/2 \Rightarrow \mathbb{E}_{S \sim D^n}[\mathcal{L}_D(A(S))] \geq \min \left\{ \frac{1}{4}, \frac{1}{2} - L\epsilon \right\} + \sigma^2,
\]

where \( M(\mathcal{X}, \epsilon, \| \cdot \|) \) is the \( \epsilon \)-packing number of \( (\mathcal{X}, \| \cdot \|) \).

Now we are ready to prove Theorem 4.1. Consider \( \mathcal{X} = \{x \in \mathbb{R}^d : ||x|| \leq 1\} \). We have \( M(\mathcal{X}, \eta, \| \cdot \|) \geq \left( \frac{1}{\eta} \right)^d \). Thus there exists a distribution \( D \) such that if \( \sigma^2 \leq 0.5 \),

\[
n < \frac{1}{2} \left( \frac{1}{\eta} \right)^d \Rightarrow n < M(\mathcal{X}, \eta, \| \cdot \|)/2 \Rightarrow \mathbb{E}_{S \sim D^n}[\mathcal{L}_D(A(S))] \geq \min \left\{ \frac{1}{4}, \frac{1}{2} - L\eta \right\} + \sigma^2.
\]

Taking \( \eta = \frac{1/2 - \epsilon}{L} \) where \( \epsilon \in (0, 1/2) \), we have \( n < \frac{1}{L} \left( \frac{2L}{1/2 - \epsilon} \right)^d \) implies \( \mathbb{E}_{S \sim D^n}[\mathcal{L}_D(A(S))] \geq \min \{ \frac{1}{4}, \epsilon \} + \sigma^2 \). Thus in the worst case, \( n \) has to be at least \( \exp(\Omega(d)) \) if one wants to achieve good astuteness by any learning algorithm that returns an \( O(1) \)-Lipschitz function. This completes the proof of Theorem 4.1.

5 Discussions

In this section, we discuss the assumptions of our results.
5.1 J-Lipschitz parametrization

We now state our results when we make an extra assumption that the parametrization of \( f \) is \( J \)-Lipschitz. The assumption was previously considered in [BS21]. We defer the detailed proof of the following theorem to Appendix C.

**Theorem 5.1.** Let \( \mathcal{F} \) be any class of functions from \( \mathbb{R}^d \to [-1, 1] \) with the \( J \)-Lipschitz parametrization with \( w \in \mathcal{W} \subseteq \mathbb{R}^p \). That is, given any norm \( || \cdot || \), for all parametrized function \( f_w \) and \( f_{w'} \) in \( \mathcal{F} \), \( || f_w - f_{w'} ||_{\mathcal{F}} \leq J || w - w' || \), where \( w, w' \in \mathcal{W} \subseteq \mathbb{R}^p \) and \( \text{diam}(\mathcal{W}) \leq W \) under the norm \( || \cdot || \). Let \( \{(x_i, y_i)\}_{i=1}^n \) be i.i.d. input-output pairs in \( \mathbb{R}^d \times [-1, 1] \). Assume that the expected conditional variance of the output (i.e., the “noise level”) is strictly positive, denoted by \( \sigma^2 := \mathbb{E}[\text{Var}[y|x]] > 0 \). Then, with high probability over the sampling of the data, one has simultaneously for all \( f \in \mathcal{F} \):

\[
\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \Rightarrow J \geq \frac{\epsilon}{3 \times 2^8 W \exp \left[ \left( \frac{\epsilon}{2^{7/2} \sqrt{\frac{n}{p}}} - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{p}} \right)^2 \right]}. 
\]

Theorem 5.1 states that with the assumption that \( \mathcal{F} \) admits a \( J \)-Lipschitz parametrization, \( J \) could be exponentially large w.r.t. \( \frac{n}{p} \), which supports our argument that we should not neglect the term \( \ln J \) in Theorem 2.1. Notice that Theorem 5.1 is also more general than Theorem 2.1, as we make no assumption on the data distribution and allow the data to be even unbounded.

5.2 Beyond label noise assumption

Our results throughout the paper hold true even without the label noise assumption. For example, in Theorem 3.1 we only need to assume the network is overfitted, i.e., \( \mathcal{L}_S(f) \leq \mathcal{L}_D(f) - \epsilon \), which leads to

\[
\mathcal{L}_S(f) \leq \mathcal{L}_D(f) - \epsilon \Rightarrow \text{Lip}_{|| \cdot ||}(f) \geq \Omega(\epsilon n^{1/d}).
\]

Notice, this assumption is natural in the practical settings, because the training error is usually less than the test error. Besides, in Theorem 4.1 we can take \( \sigma = 0 \) to get ride of the label noise assumption.

6 Conclusions

In this work, we study the sample complexity of robust interpolation problem when the data are in a unit ball. We show that both too many data and small data hurt robustness. Our analysis reveals \( \exp(\Theta(d)) \) as the sample complexity of robust interpolation problem for certain distributions, though benign data distributions may lead to improved sample complexity. Perhaps surprisingly, the results shed light on the curse of big data and the blessing of dimensionality regarding robustness, and discover an intriguing phase transition phenomenon of astuteness at \( n = \exp(\Theta(d)) \).

References

[BCM19] Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems*, 2019.

[BDMZ20] Avrim Blum, Travis Dick, Naren Manoj, and Hongyang Zhang. Random smoothing might be unable to certify \( \ell_\infty \) robustness for high-dimensional images. *Journal of Machine Learning Research*, 21:1–21, 2020.
[BJC21] Robi Bhattacharjee, Somesh Jha, and Kamalika Chaudhuri. Sample complexity of robust linear classification on separated data. In International Conference on Machine Learning, pages 884–893, 2021.

[BLN21] Sebastien Bubeck, Yuanzhi Li, and Dheeraj M Nagaraj. A law of robustness for two-layers neural networks. In Annual Conference on Learning Theory, volume 134, pages 804–820, 2021.

[BS21] Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. In Advances in Neural Information Processing Systems, 2021.

[BTEGN09] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. Robust optimization. Princeton university press, 2009.

[CBM18] Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. PAC-learning in the presence of evasion adversaries. In Advances in Neural Information Processing Systems, pages 230–241, 2018.

[DHHR20] Edgar Dobriban, Hamed Hassani, David Hong, and Alexander Robey. Provable tradeoffs in adversarially robust classification. arXiv preprint arXiv:2006.05161, 2020.

[DWR20] Chen Dan, Yuting Wei, and Pradeep Ravikumar. Sharp statistical guarantees for adversarially robust gaussian classification. In International Conference on Machine Learning, pages 2345–2355, 2020.

[GCL+19] Ruiqi Gao, Tianle Cai, Haochuan Li, Cho-Jui Hsieh, Liwei Wang, and Jason D Lee. Convergence of adversarial training in overparametrized neural networks. Advances in Neural Information Processing Systems, 32:13029–13040, 2019.

[GSS14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In International Conference on Learning Representations, 2014.

[HJ22] Hamed Hassani and Adel Javanmard. The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression. arXiv preprint arXiv:2201.05149, 2022.

[Hub04] Peter J Huber. Robust statistics, volume 523. John Wiley & Sons, 2004.

[KLGF20] Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. In International Conference on Machine Learning, pages 5458–5467, 2020.

[KT61] Andrei N Kolmogorov and Vladimir M Tihomirov. $\epsilon$-entropy and $\epsilon$-capacity of sets in functional spaces. Mathematics and Its Applications, 17:277–364, 1961.

[MHS19] Omar Montasser, Steve Hanneke, and Nathan Srebro. VC classes are adversarially robustly learnable, but only improperly. In Annual Conference on Learning Theory, pages 2512–2530, 2019.

[MMS+17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations, 2017.

[MV03] Shahar Mendelson and Roman Vershynin. Entropy and the combinatorial dimension. Inventiones mathematicae, 152(1):37–55, 2003.
[NAM21] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In NeurIPS 2021 Datasets and Benchmarks Track, 2021.

[RWK20] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In International Conference on Machine Learning, pages 8093–8104, 2020.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.

[SST+18] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In Advances in Neural Information Processing Systems, 2018.

[SYS+19] Hadi Salman, Greg Yang, Jerry Li, Pengchuan Zhang, Huan Zhang, Ilya Razenshteyn, and Sebastien Bubeck. Provably robust deep learning via adversarially trained smoothed classifiers. In Advances in Neural Information Processing Systems, 2019.

[SZS+14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In International Conference on Learning Representations, 2014.

[vLB04] Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with Lipschitz functions. Journal of Machine Learning Research, 5:669–695, 2004.

[WJC18] Yizhen Wang, Somesh Jha, and Kamalika Chaudhuri. Analyzing the robustness of nearest neighbors to adversarial examples. In International Conference on Machine Learning, pages 5133–5142. PMLR, 2018.

[YDH+20] Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In International Conference on Machine Learning, pages 10693–10705, 2020.

[YKB19] Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In International conference on machine learning, pages 7085–7094, 2019.

[YZR+20] Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Ruslan Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. In Advances in Neural Information Processing Systems, 2020.

[ZYJ+19] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In International Conference on Machine Learning, pages 7472–7482, 2019.
A Proof of Theorem 3.1

Definition A.1 (Representativeness of $S$).

\[ \text{Rep}_D(l, \mathcal{F}, S) := \sup_{f \in \mathcal{F}} (\mathcal{L}_D(f) - \mathcal{L}_S(f)). \]

Definition A.2 (Rademacher complexity). For $A \in \mathbb{R}^p$,

\[ R(A) := \frac{1}{n} \mathbb{E}_{\sigma_1, \ldots, \sigma_n \in \{-1, 1\}} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right]. \]

Lemma A.3. Assume that $\forall f \in \mathcal{F}, \forall x \in \mathcal{X}, |l(f, x)| \leq c$. Then with probability at least $1 - \delta$, for all $f \in \mathcal{F}$,

\[ \mathcal{L}_D(f) - \mathcal{L}_S(f) \leq \mathbb{E}_{S \in D^n}[\text{Rep}_D(l, \mathcal{F}, S)] + c\sqrt{\frac{2\ln(2/\delta)}{n}}. \]

Lemma A.4 (Lemma 26.2 in [SSBD14]).

\[ \mathbb{E}_{S \in D^n}[\text{Rep}_D(l, \mathcal{F}, S)] \leq 2\mathbb{E}_{S \in D^n}[R(l \circ \mathcal{F} \circ S)], \]

where $S = \{x_1, \ldots, x_n\}$ and $l \circ \mathcal{F} \circ S = \{(l(f, x_1, y_1), \ldots, l(f, x_n, y_n)) \in \mathbb{R}^p\}$.

Lemma A.5 (Theorem 26.5 in [SSBD14]). Assume $\forall f \in \mathcal{F}, \forall x \in \mathcal{X}, |l(f, x)| \leq a$, then with probability at least $1 - \delta$, for all $f \in \mathcal{F}$,

\[ \mathcal{L}_D(f) - \mathcal{L}_S(f) \leq 2\mathbb{E}_{S' \in D^n}[R(l \circ \mathcal{F} \circ S')] + a\sqrt{\frac{2\ln(2/\delta)}{n}}. \]

Lemma A.6 (Lemma 26.9 in [SSBD14]). If $l(f(x), y)$ is $C_{\|\cdot\|}$-Lipschitz w.r.t. $f(x)$ for arbitrary $y \in [-1, 1]$,

\[ R(l \circ \mathcal{F} \circ S) \leq C \cdot R(\mathcal{F} \circ S). \]

Definition A.7 (Covering number). For a given space $(\mathcal{X}, \|\cdot\|)$ and a covering radius $\eta$, $N(\mathcal{X}, \eta, \|\cdot\|)$ is the minimum number of $\eta$-ball which cover the whole space of $\mathcal{X}$. Note the metric $\|\cdot\|_{\mathcal{F}}$ on the function space $\mathcal{F}$ is

\[ \|f - f'\|_{\mathcal{F}} = \sup_x |f(x) - f'(x)|. \]

Lemma A.8. For a bounded space $(\mathcal{X}, \|\cdot\|)$ and $B_L := \{f : \text{Lip}_{\|\cdot\|}(f) \leq L\}$, if $\mathcal{X}$ is connected and centered, we have for every $\epsilon > 0$

\[ N(B_L, \epsilon, \|\cdot\|) \leq \left( \left\lceil \frac{2L \cdot \text{diam}(\mathcal{X})}{\epsilon} \right\rceil \right) 2^{N(\mathcal{X}, \sqrt{\pi \epsilon}, \|\cdot\|)}. \]

Proof. We consider the Lipschitz function class $B_L := \{f : \text{Lip}_{\|\cdot\|}(f) \leq L\}$. In order to bound the covering number of $\mathcal{F}$, we consider an $\frac{\epsilon}{\pi \cdot \text{diam}(\mathcal{X})}$-covering of input space $\mathcal{X}$ consisting of $N = N_{\epsilon/(2L)}(\mathcal{X})$ plates $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_N$ centered at $s_1, s_2, \ldots, s_N$. The fact that $\mathcal{X}$ is connected enables one to join any two sets $\mathcal{U}_i$ and $\mathcal{U}_j$ by a chain of intersecting $\mathcal{U}_k$. For any function $f \in \mathcal{F}$, we can construct its approximating functional $\tilde{f}$ by taking its value on $\mathcal{U}_1$ as an $\epsilon/2$-approximation of $f(s_1)$. As $\text{diam}(\mathcal{U}_1) \leq L \cdot \text{diam}(\mathcal{X})$, there are at most $\lceil 2L \cdot \text{diam}(\mathcal{X})/\epsilon \rceil$ such approximations. On the other hand, note that the $N$ plates are chained. By Lipschitzness, the function values of $f$ on $s_1$ and $s_2$ differ at most $\epsilon/2$, and so $f(s_2)$ differs at most $\epsilon$ from $\tilde{f}(s_1)$ by triangle inequality. It implies that to construct an $\epsilon$-approximation of $f(s_2)$ on $\mathcal{U}_2$, we shall know either $\tilde{f}(s_1) - \epsilon/2$ or $\tilde{f}(s_1) + \epsilon/2$. Repeating the same argument by $N$ times, we can bound the $\epsilon$-covering of $f$ on $\mathcal{X}$ by $[2L \cdot \text{diam}(\mathcal{X})/\epsilon]2^N$. \qed
Theorem A.9 (Theorem 18 of [vLB04]). Let \((\mathcal{X}, \| \cdot \|)\) be a bounded space and \(B_L\) be the set of functions \(f\)'s such that \(\text{Lip}_{\| \cdot \|}(f) \leq L\), \(n = |S|\). If \(\mathcal{X}\) is connected and centered, for any \(\epsilon > 0\)

\[
\mathbb{E}_{S' \in \mathcal{D}^n}[R(B_L \circ S')] \leq 2\epsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\epsilon/4}^{2L \cdot \text{diam}(\mathcal{X})} \sqrt{N\left(\mathcal{X}, \frac{u}{2L}, \| \cdot \|\right)} \ln 2 + \ln \left( \left\lceil \frac{2L \cdot \text{diam}(\mathcal{X})}{u} \right\rceil \right) du.
\]

Lemma A.10 (Equation 1 of [MV03]). For any \(\mathcal{X}\) with finite diameter w.r.t. \(\| \cdot \|\), if \(\epsilon \leq \text{diam}(\mathcal{X})\),

\[
N(\mathcal{X}, \epsilon, \| \cdot \|) \leq \left( \frac{3\text{diam}(\mathcal{X})}{\epsilon} \right)^d.
\]

Corollary A.11. If \(\text{diam}(\mathcal{X}) = 2\) w.r.t. \(\| \cdot \|\), we have

\[
\mathbb{E}_{S \in \mathcal{D}^n}[R(B_L \circ S)] = \mathcal{O}\left( \frac{L}{n^{1/d}} \right).
\]

**Proof.** As \(\frac{\epsilon}{2L} \leq \text{diam}(\mathcal{X})\), we have \(N(\mathcal{X}, \frac{\epsilon}{2L}, \| \cdot \|) \leq \left( \frac{12L}{\epsilon} \right)^d\) and

\[
\mathbb{E}_{S \in \mathcal{D}^n}[R(B_L \circ S)] \leq 2\epsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\epsilon/4}^{4L} \sqrt{\left( \frac{12L}{u} \right)^d} \ln 2 + \ln \left( \left\lceil \frac{2L}{u} \right\rceil \right) du
\]

\[
\leq 2\epsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\epsilon/4}^{4L} \left[ \sqrt{\left( \frac{12L}{u} \right)^d} \ln 2 + \sqrt{\ln \left( \left\lceil \frac{2L}{u} \right\rceil \right)} \right] du
\]

\[
\leq 2\epsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\epsilon/4}^{4L} \sqrt{\left( \frac{12L}{u} \right)^d} \ln 2 du + \frac{16\sqrt{2}L}{\sqrt{n}} \sqrt{\ln(16L/\epsilon + 1)}.
\]

Switching the integral variable from \(u\) to \(v = u/12L\) we have

\[
\int_{\epsilon/4}^{4L} \sqrt{\left( \frac{12L}{u} \right)^d} \ln 2 du = 12L \int_{\epsilon/(48L)}^{1/3} \sqrt{v^{-d}} \ln 2 dv
\]

\[
= 12L \left[ \sqrt{\ln 2} \cdot \frac{1}{-d/2 + 1} v^{-d/2 + 1/3} \bigg|_{\epsilon/(48L)}^{1/3} \right]
\]

\[
< 12L \frac{2\sqrt{\ln 2}}{d - 2} \left( \frac{48L}{\epsilon} \right)^{d/2 - 1}.
\]

Based on the calculation above we have

\[
\mathbb{E}_{S \in \mathcal{D}^n}[R(B_L \circ S)] \leq 2\epsilon + L \frac{96\sqrt{2} \ln 2}{\sqrt{n}(d - 2)} \left( \frac{48L}{\epsilon} \right)^{d/2 - 1} + \frac{16\sqrt{2}L}{\sqrt{n}} \sqrt{\ln(16L/\epsilon + 1)}.
\]

As this inequality holds for arbitrary \(\epsilon > 0\), we can take \(\epsilon = 48L/n^{1/d}\) and have

\[
\mathbb{E}_{S \in \mathcal{D}^n}[R(B_L \circ S)] \leq 96 \frac{L}{n^{1/d}} + \frac{96\sqrt{2} \ln 2}{d - 2} \frac{L}{n^{1/d}} + \frac{16\sqrt{2}L}{\sqrt{n}} \sqrt{\ln \left( \frac{1}{3} n^{1/d} + 1 \right)} \sim \mathcal{O}\left( \frac{L}{n^{1/d}} \right).
\]
Lemma A.12. If \( l(f(x), y) = (f(x) - y)^2 \), \( l \) is 4-Lipschitz continuous w.r.t. \( f(x) \) and
\[
\mathbb{E}_{S \in D^n}[R(l \circ B_L \circ S)] \leq 4 \mathbb{E}_{S \in D^n}[R(B_L \circ S)] = O \left( \frac{L}{n^{1/d}} \right).
\]

Theorem A.13. Let \( \mathcal{X} \subseteq \mathbb{R}^d \), \( \text{diam}(\mathcal{X}) = 2 \) and the function space \( \mathcal{F} = \{ f : \mathcal{X} \to [-1, 1] \} \). Assume that the observation noise has variance \( \sigma^2 = \mathbb{E}_x \text{Var}(y|x) \). We have with probability \( 1 - \delta \), for every \( f \in \mathcal{F} \):
\[
\mathcal{L}_S(f) \leq \sigma^2 - \epsilon \Rightarrow \text{Lip}_{\| \cdot \|}(f) \geq \Omega(\epsilon n^{1/d}).
\]

Proof. According to Lemma A.5,
\[
\mathcal{L}_D(f) - \mathcal{L}_S(f) \leq 2 \mathbb{E}_{S \in D^n}[R(l \circ F \circ S)] + a \sqrt{\frac{2 \ln(2/\delta)}{n}}.
\]
where \( a = \max_{(x,y)} l(f(x), y) \leq 4 \). According to Lemma A.6 and \( \nabla f(x) \leq 4 \),
\[
\mathbb{E}_{S \in D^n}[R(l \circ F \circ S)] \leq 4 \mathbb{E}_{S \in D^n}[R(F \circ S)].
\]
Thus we have
\[
\mathbb{E}_{S \in D^n}[R(F \circ S)] \geq \frac{1}{8} \left( \mathcal{L}_D(f) - \mathcal{L}_S(f) - 4 \sqrt{\frac{2 \ln(2/\delta)}{n}} \right),
\]
Under the noisy label settings, we have
\[
\mathcal{L}_D(f) = \mathbb{E}_D[(f(x) - y)^2] = \mathbb{E}_{x,y}[(f(x) - \mathbb{E}_y[y|x])^2 + (y - \mathbb{E}_y[y|x])^2] \geq \mathbb{E}_x[\text{Var}(y|x)] = \sigma^2.
\]
So with the overfitting assumption \( \mathcal{L}_S(f) \leq \sigma^2 - \epsilon \), we have
\[
\mathbb{E}_{S \in D^n}[R(F \circ S)] \geq \frac{1}{8} \left( \mathcal{L}_D(f) - \mathcal{L}_S(f) - 4 \sqrt{\frac{2 \ln(2/\delta)}{n}} \right) \geq \frac{1}{8} \left( \sigma^2 - \sigma^2 + \epsilon - 4 \sqrt{\frac{2 \ln(2/\delta)}{n}} \right) = \frac{\epsilon}{8} - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{n}}.
\]
Consider \( B_L = \{ f \in \mathcal{F} : \text{Lip}_{\| \cdot \|}(f) \leq L \} \). According to Corollary A.11, we have
\[
K \frac{L}{n^{1/d}} \geq \mathbb{E}_{S' \in D^n}[R(B_L \circ S')] \geq \frac{\epsilon}{8} - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{n}},
\]
where \( K = 96 + \frac{96\sqrt{2} \ln 2}{d-2} + \frac{16\sqrt{2}}{n^{1/2-1/d}} \sqrt{\ln(\frac{1}{3} n^{1/d} + 1)} \sim \Theta(1) \). Thus we have
\[
L \geq Kn^{1/d} \left( \frac{1}{8} \epsilon - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{n}} \right) \sim \Omega(\epsilon n^{1/d}).
\]
If \( \exists f_0 \in \mathcal{F} \), such that
\[
\mathcal{L}_S(f_0) \leq \sigma^2 - \epsilon \Rightarrow \text{Lip}_{\| \cdot \|}(f_0) < \frac{n^{1/d}}{K} \left( \frac{1}{8} \epsilon - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{n}} \right),
\]
we have
\[
\frac{\epsilon}{8} - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{n}} \geq K \frac{\text{Lip}_{\| \cdot \|}(f_0)}{n^{1/d}} \geq
\]
\[
\mathbb{E}_{S \in \mathcal{D}^n}[R(B_{\text{Lip}_{\| \cdot \|}(f_0)} \circ S)] \geq \frac{\epsilon}{8} - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{n}},
\]
which yields contradiction. Therefore \( \forall f \in \mathcal{F} \),
\[
\mathcal{L}_S(f) \leq \sigma^2 - \epsilon \Rightarrow \text{Lip}_{\| \cdot \|}(f) \geq \frac{n^{1/d}}{K} \left( \frac{1}{8} \epsilon - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{n}} \right).
\]
\[\square\]

## B Proof of Theorem 4.1

### Lemma B.1.
Let \( \mathcal{A}(S) : \mathcal{X} \to \{-a,a\} \) be any learning algorithm w.r.t. the squared \( \ell_2 \) loss over a domain \( \mathcal{X} \) and samples \( S \). Assume there are label noise \( \mathbb{E} [\text{Var}[y|x]] = \sigma^2 \). Let \( m \) be any number smaller than \( |\mathcal{X}|/2 \), representing a training set size. Then, for any \( a > 0 \) there exists a distribution \( \mathcal{D} \) (with label noise) over \( \mathcal{X} \times \{-a,a\} \) such that \( \mathbb{E}_{S \sim \mathcal{D}^n}[\mathcal{L}_D(\mathcal{A}(S))] \geq \frac{1}{2} (a^2 + \sigma^2) \).

**Proof.** Our proof is partly based on Theorem 5.1 of [SSBD14]. Let \( C \) be a subset of \( \mathcal{X} \) of size \( 2m \). There exist \( T = 2^{2m} \) possible labeling functions from \( C \) to \( \{-a,a\} \). Denote these functions by \( f_1, \ldots, f_T \). We then define a distribution \( \mathcal{D}_i \) w.r.t. \( f_i \) by
\[
\mathcal{D}_i(\{(x,y)\}) = \begin{cases} p/|C|, & \text{if } y = f_i(x); \\ (1-p)/|C|, & \text{if } y \neq f_i(x), \end{cases}
\]
where \( p > 1/2 \) satisfies \( \text{Var}(y|x) = \sigma^2 = 4a^2 p(1-p) \) (notice that as \( f_i(x) \) can only be \( a \) or \(-a\), \( p \) is the same for all \( f_i(x) \)'s). In this way, \( \mathcal{D}_i \) satisfies the noisy label setting. We will show that for every algorithm \( \mathcal{A} \) that receives a training set of size \( m \) from \( C \times \{-a,a\} \) and returns a function \( \mathcal{A}(S) : C \to \mathbb{R} \), it holds that
\[
\max_{i \in [T]} \mathbb{E}_{S \sim \mathcal{D}_i^n}[\mathcal{L}_{\mathcal{D}_i}(\mathcal{A}(S))] \geq \frac{a^2 + \sigma^2}{2}.
\]
There are \( k = (2m)^m \) possible sequences of \( m \) instances from \( C \). Denote these sequences by \( S_1, \ldots, S_k \). Also, if \( S_j = (x_1, \ldots, x_m) \), we denote by \( S_j^i \) the sequence containing the instances in \( S_j \) labeled by the function \( f_i \), namely, \( S_j^i = ((x_1, a_1 f_i(x_1)), \ldots, (x_m, a_m f_i(x_m))) \), where \( \mathbb{P}(a_i = 1) = p \), \( \mathbb{P}(a_i = -1) = 1 - p \), and \( a_1, \ldots, a_m \) are i.i.d. for all \( S_j^i \), given that \( p \) is the same for all \( f_i(x) \)'s. If the distribution is \( \mathcal{D}_i \), then the possible training sets that algorithm \( \mathcal{A} \) receives are \( S_j^1, \ldots, S_j^k \), and all these training sets have the same probability of being sampled. Therefore,
\[
\mathbb{E}_{S \sim \mathcal{D}_i^n}[\mathcal{L}_{\mathcal{D}_i}(\mathcal{A}(S))] = \frac{1}{k} \sum_{j=1}^k \mathcal{L}_{\mathcal{D}_i}(\mathcal{A}(S_j^i)).
\]
Using the facts that “maximum” is larger than “average” and that “average” is larger than “minimum”, we have

\[
\max_{i \in [T]} \frac{1}{k} \sum_{j=1}^{k} \mathcal{L}_{D_i}(A(S^i_j)) \geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{k} \sum_{j=1}^{k} \mathcal{L}_{D_i}(A(S^i_j)) = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{T} \sum_{i=1}^{T} \mathcal{L}_{D_i}(A(S^i_j)) \geq \min_{j \in [k]} \frac{1}{T} \sum_{i=1}^{T} \mathcal{L}_{D_i}(A(S^i_j)).
\]

Next, fix some \( j \in [k] \). Denote by \( S_j := (x_1, \ldots, x_m) \) and let \( v_1, \ldots, v_q \) be the instances in \( C \) that do not appear in \( S_j \). Clearly, \( q \geq m \). Therefore, for every function \( h : C \to \mathbb{R} \) and every \( i \) we have

\[
\mathcal{L}_{D_i}(h) = \frac{1}{2m} \mathbb{E}_{a \in \{-1,1\}^{2m}} \left[ \sum_{x \in C} (h(x) - a_i f_i(x))^2 \right] = \frac{1}{2m} \sum_{x \in C} \left[ p(h(x) - f_i(x))^2 + (1-p)(h(x) + f_i(x))^2 \right] = \frac{1}{2m} \sum_{x \in C} [(h(x) - (2p-1)f_i(x))^2 + 4p(1-p)f_i(x)^2] = \sigma^2 + \frac{1}{2m} \sum_{x \in C} [(h(x) - (2p-1)f_i(x))^2].
\]

Note that

\[
\frac{1}{2m} \sum_{x \in C} [(h(x) - (2p-1)f_i(x))^2] \geq \frac{1}{2m} \sum_{r=1}^{q} (h(v_r) - (2p-1)f_i(v_r))^2 \geq \frac{1}{2q} \sum_{r=1}^{q} (h(v_r) - (2p-1)f_i(v_r))^2.
\]

Hence,

\[
\frac{1}{T} \sum_{i=1}^{T} \mathcal{L}_{D_i}(A(S^i_j)) \geq \frac{1}{T} \sum_{i=1}^{T} \mathbb{E}_{a \in \{-1,1\}^{m}} \left[ \sigma^2 + \frac{1}{2q} \sum_{r=1}^{q} (A(S^i_j(a))(v_r) - (2p-1)f_i(v_r))^2 \right] = \sigma^2 + \frac{1}{2q} \sum_{r=1}^{q} \frac{1}{T} \sum_{i=1}^{T} \mathbb{E}_{a \in \{-1,1\}^{m}} [(A(S^i_j(a))(v_r) - (2p-1)f_i(v_r))^2] \geq \sigma^2 + \frac{1}{2q} \min_{r \in [p]} \frac{1}{T} \sum_{i=1}^{T} \mathbb{E}_{a \in \{-1,1\}^{m}} [(A(S^i_j(a))(v_r) - (2p-1)f_i(v_r))^2].
\]

Next, fix some \( r \in [p] \). We can partition all the functions in \( f_1, \ldots, f_T \) into \( T/2 \) disjoint pairs, where for a pair \((f_i, f_{r'})\) we have that for every \( c \in C, f_i(c) \neq f_{r'}(c) \) if and only if \( c = v_r \). Note that for such a pair and the same \( a \), we must have \( S^i_j(a) = S^j_{r'}(a) \) and \( \forall \alpha \in \{-1,1\}^m, \mathbb{P}(a|S_j^i) = \mathbb{P}(a|S_{r'}^j) \). It follows that

\[
\mathbb{E}_{a \in \{-1,1\}^{m}} [(A(S^i_j(a))(v_r) - (2p-1)f_i(v_r))^2] \geq \mathbb{E}_{a \in \{-1,1\}^{m}} [(A(S^j_{r'}(a))(v_r) - (2p-1)f_i(v_r))^2] \geq \mathbb{E}_{a \in \{-1,1\}^{m}} \left[ \frac{1}{2}(2p-1)^2(f_{r'}(v_r) - f_i(v_r))^2 \right] = 2(2p-1)^2 \sigma^2,
\]

\[
= 2(2p-1)^2 \sigma^2.
\]
which yields

\[
\frac{1}{T} \sum_{i=1}^{T} \mathbb{E}_{a \in \{-1,1\}^n} [(A(S_i^j)(a))(v_r) - (2p - 1)f_i(v_r))^2] \geq (2p - 1)^2a^2.
\]

Combining the discussion above, we have

\[
\max_{i \in [T]} \mathbb{E}_{S \sim \mathcal{D}^n}[\mathcal{L}_{\mathcal{D}}(A(S))] \geq \min_{j \in [k]} \frac{1}{T} \sum_{i=1}^{T} \mathcal{L}_{\mathcal{D}}(A(S_i^j)) \geq \sigma^2 + \frac{1}{2}(2p - 1)^2a^2 = \frac{a^2 + \sigma^2}{2}.
\]

\[\square\]

**Theorem B.2** (No-free-lunch theorem for Lipschitz learning algorithm). Let \(A(S) : \mathcal{X} \to \mathbb{R}\) be any \(L\)-Lipschitz (w.r.t. norm \(\| \cdot \|\)) learning algorithm for the task of regression w.r.t. the squared \(\ell_2\) loss over a domain \((\mathcal{X}, \| \cdot \|)\) and samples \(S\). Let \(n\) be the size of training set. Assume observation noise \(\sigma^2 = \mathbb{E}_D[\text{Var}(y|x)] \leq \frac{1}{2}\). Then, there exists a distribution \(\mathcal{D}\) over \(\mathcal{X} \times [-1, 1]\) such that for all \(L\)-Lipschitz (w.r.t. norm \(\| \cdot \|\)) learning algorithm and any \(\epsilon \in [0, \frac{1}{3}]\):

\[
n < M(\mathcal{X}, \epsilon, \| \cdot \|)/2 \Rightarrow \mathbb{E}_{S \sim \mathcal{D}^n}[\mathcal{L}_{\mathcal{D}}(A(S))] \geq \min \left\{ \frac{1}{4}, \frac{1}{2} - L\epsilon \right\} + \sigma^2,
\]

where \(M(\mathcal{X}, \epsilon, \| \cdot \|)\) is the \(\epsilon\)-packing number of space \((\mathcal{X}, \| \cdot \|)\).

**Proof.** Consider an arbitrary finite set \(\mathcal{C} \subseteq \mathcal{X}\). Denote by \(d(\mathcal{C}) := \min_{(a,b) \in \mathcal{C} \times \mathcal{C}, a \neq b} \|a - b\|\). We now consider two cases: a) \(d(\mathcal{C}) < \epsilon\) and b) \(d(\mathcal{C}) \geq \epsilon\), and show that our conclusion holds for both cases.

**Case a):** \(d(\mathcal{C}) < \epsilon\). Denote by \((x_1, x_2) = \arg \min_{(a,b) \in \mathcal{C} \times \mathcal{C}, a \neq b} \|a - b\|\). We can select \(\mathcal{D}\) such that \(\mathcal{D}((x_1, 1)) = \frac{p}{2}, \mathcal{D}((x_2, -1)) = \frac{1-p}{2}\). Consider an \(L\)-Lipschitz learning algorithm \(A(S) : \mathcal{C} \to \mathbb{R}\):

\[
\mathbb{E}_{S \sim \mathcal{D}^n}[\mathcal{L}_{\mathcal{D}}(A(S))] \cong \min_{S \sim \mathcal{D}^n} \left[ \frac{\mathcal{P}}{2}(A(S)(x_1) - 1)^2 + \frac{1-p}{2}(A(S)(x_1) + 1)^2 + \frac{p}{2}(A(S)(x_2) + 1)^2 + \frac{1-p}{2}(A(S)(x_2) - 1)^2 \right] \cong \min_{S \sim \mathcal{D}^n} \left[ 1 - (2p - 1)|A(S)(x_1) - A(S)(x_2)| \right] \geq 1 - \mathcal{L}(2p - 1)|x_1 - x_2| \geq 1 - L \cdot d(\mathcal{C}) = 1 - L\epsilon \geq \frac{1}{2} - L\epsilon + \sigma^2.
\]

**Case b):** \(d(\mathcal{C}) \geq \epsilon\). We reduce the regression problem from a binary classification problem with target \(\{-1, 1\}\) by considering the distribution \(\mathcal{D}\) such that \(\mathcal{D}\) only on \(\mathcal{X} \times \{-1, 1\}\). Then by **Theorem B.1**, for every \(A(S) : \mathcal{X} \to \mathbb{R}\) and every \(\mathcal{C} \subseteq \mathcal{X}\) there exists \(\mathcal{D}\) such that

\[
n < \frac{|\mathcal{C}|}{2} \Rightarrow \mathbb{E}_{S \sim \mathcal{D}^n}[\mathcal{L}_{\mathcal{D}}(A(S))] \geq \frac{1 + \sigma^2}{2}.
\]

Notice that \(\mathcal{C} \subseteq \mathcal{X}\) can be chosen arbitrarily. Thus we have

\[
n < \max_{\mathcal{C} \subseteq \mathcal{X}, d(\mathcal{C}) \geq \epsilon} \frac{|\mathcal{C}|}{2} \Rightarrow \mathbb{E}_{S \sim \mathcal{D}^n}[\mathcal{L}_{\mathcal{D}}(A(S))] \geq \frac{1 + \sigma^2}{2}.
\]
Denote the $\epsilon$-packing number of space $(\mathcal{X}, || \cdot ||)$ by $M(\mathcal{X}, \epsilon, || \cdot ||)$. We have

$$\max_{C \subseteq X, d(c) \geq \epsilon} \frac{|C|}{2} = \frac{M(\mathcal{X}, \epsilon, || \cdot ||)}{2}.$$ 

That is,

$$n < M(\mathcal{X}, \epsilon, || \cdot ||)/2 \Rightarrow \mathbb{E}_{S \sim \mathcal{D}^n}[\mathcal{L}_\mathcal{D}(A(S))] \geq \frac{1 + \sigma^2}{2} \geq \frac{1}{4} + \sigma^2.$$ 

Combining a) and b) yields our conclusion.

**Theorem B.3.** Let $S = \{(x_i, y_i)\}_{i=1}^n$ be i.i.d. training pairs in $\{x : ||x|| \leq 1\} \times [-1, 1]$ for any given norm $|| \cdot ||$. Denote by $\mathcal{L}_\mathcal{D}(f) := \mathbb{E}_D[(f(x) - y)^2]$ the squared $\ell_2$ loss. Assume that the expected conditional variance of the output (i.e., the “noise level”) is strictly positive and bounded by $\frac{\sigma^2}{2}$, denoted by $\sigma^2 := \mathbb{E}[\text{Var}(y|x)]$. Let $A(S) : \mathcal{X} \to \mathbb{R}$ be any $L$-Lipschitz learning algorithm over a training set $S$. Then there exists a distribution $\mathcal{D}'$ of $(x, y)$ such that

$$n < \frac{1}{2} \left( \frac{2L}{1-2\epsilon} \right)^d \Rightarrow \mathbb{E}_S[\mathcal{L}_{\mathcal{D}'}(A(S))] \geq \min \left\{ \frac{1}{4}, \epsilon \right\} + \sigma^2.$$ 

**Proof.** Consider $\mathcal{X} = \{x \in \mathbb{R}^d : ||x|| \leq 1\}$. We have $M(\mathcal{X}, \eta, || \cdot ||) \geq \left( \frac{1}{\eta} \right)^d$ and thus there exists a distribution $\mathcal{D}$ such that if $\sigma^2 \leq 0.5$

$$n < \frac{1}{2} \left( \frac{1}{\eta} \right)^d \Rightarrow n < M(\mathcal{X}, \eta, || \cdot ||)/2 \Rightarrow \mathbb{E}_{S \sim \mathcal{D}^n}[\mathcal{L}_\mathcal{D}(A(S))] \geq \min \left\{ \frac{1}{4}, \frac{1}{2} - L\eta \right\} + \sigma^2.$$ 

Taking $\eta = \frac{1/2 - \epsilon}{L}$ where $\epsilon \in (0, 1/2)$, we have

$$n < \frac{1}{2} \left( \frac{2L}{1-2\epsilon} \right)^d \Rightarrow \mathbb{E}_{S \sim \mathcal{D}^n}[\mathcal{L}_\mathcal{D}(A(S))] \geq \min \left\{ \frac{1}{4}, \epsilon \right\} + \sigma^2.$$ 

Thus in the worst case, $n$ has to be at least $\exp(\Omega(d))$ if one wants to achieve good astuteness by any $O(1)$-Lipschitz learning algorithm, this completes our proof.

**C Proof of Theorem 5.1**

**Theorem C.1.** Let $\mathcal{F}$ be any class of functions from $\mathbb{R}^d \to [-1, 1]$ with the $J$-Lipschitz parametrization with $w \in \mathcal{W} \subseteq \mathbb{R}^p$. That is, given a norm $|| \cdot ||$, for all parametrized function $f_w$ and $f_{w'}$ in $\mathcal{F}$, $||f_w - f_{w'}||_F \leq J||w - w'||$, where $w, w' \in \mathcal{W} \subseteq \mathbb{R}^p$ and $\text{diam}(\mathcal{W}) \leq W$ under the norm $|| \cdot ||$. Let $\{(x_i, y_i)\}_{i=1}^n$ be i.i.d. input-output pairs in $\mathbb{R}^d \times [-1, 1]$. Assume that the expected conditional variance of the output (i.e., the “noise level”) is strictly positive, denoted by $\sigma^2 := \mathbb{E}[\text{Var}(y|x)] > 0$. Then, with high probability over the sampling of the data, one has simultaneously for all $f \in \mathcal{F}$:

$$\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \Rightarrow J \geq \frac{\epsilon}{3 \times 2^{8W}} \exp \left[ \left( \frac{\epsilon}{2^7 \sqrt{2}} \sqrt{\frac{n}{p}} - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{p}} \right)^2 \right].$$

**Proof.** We argue that the $\eta$-covering of the function space $\mathcal{F}$ is upper bounded by the $\eta/J$-covering of the parameter space $\mathcal{W}$. To see this, we can select the centers $\mathcal{W}^c = \{w^c\}$ of the $\eta/J$-covering of $\mathcal{W}$, and covering $\mathcal{F}$ with $\eta$-balls centered at $f_{w^c}$, because $\forall f \in \mathcal{F}$, we can find $w' \in \mathcal{W}^c$ such that $||w - w'|| \leq \eta/J$, by the
definition of $J$-Lipschitz parametrization we have $||f_w - f_{w'}||_{\mathcal{F}} \leq J ||w - w'|| \leq \eta$, thus $\mathcal{F}$ can be covered by $N(W, \eta/J, ||\cdot||)$ balls. So we have

$$N(\mathcal{F}, \eta, ||\cdot||) \leq N(\mathcal{W}, \eta/J, ||\cdot||) \leq (6JW/\eta)^p.$$ 

According to Lemma 3.4, given $n$ training samples $S$, we have $\text{diam}(\mathcal{F}) \leq 2$ as $f \in \mathcal{F}$ is bounded by $[-1, 1]$. Thus

$$\mathbb{E}_{S' \in D^n}[R(\mathcal{F} \circ S)] \leq 2\eta + 4\sqrt{2} \int_{\eta/4}^{2} \sqrt{\ln(6JW/\eta)} du \leq 2\eta + 8\sqrt{2} \frac{\sqrt{p \ln(24JW/\eta)}}{\sqrt{n}}.$$ 

By Equation 3, we have

$$2\eta + \frac{8\sqrt{2}}{\sqrt{n}} \sqrt{p \ln(24JW/\eta)} \geq \mathbb{E}_{S' \in D^n}[R(\mathcal{F} \circ S)] \geq \epsilon/8 - \frac{1}{2} \sqrt{\frac{2 \ln(2/\delta)}{n}}.$$ 

Taking $\eta = \frac{\epsilon}{32}$ yields

$$J \geq \frac{\epsilon}{3 \times 2^6W} \exp \left[ \left( \frac{\epsilon}{2^7 \sqrt{2}} \sqrt{\frac{n}{p}} - \frac{1}{2} \sqrt{\frac{2 \ln(2\delta)}{p}} \right)^2 \right] \sim \Omega \left( \frac{\epsilon}{W} \exp \left( \frac{\epsilon^2 n}{p} \right) \right),$$

and thus

$$\mathcal{L}_S(f) \leq \sigma^2 - \epsilon \Rightarrow J \geq \Omega \left( \frac{\epsilon}{W} \exp \left( \frac{\epsilon^2 n}{p} \right) \right).$$

\[\square\]

**D Overparametrization is Necessary for Good Data Fitting**

**Lemma D.1.** If $a \leq X \leq b$ with probability 1, then $X$ is a $\frac{(b-a)^2}{4}$-subgaussian random variable.

**Lemma D.2.** If $X_1, \ldots, X_n$ are independent, $C$-subgaussian with mean 0, then $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$ is $18C$-subgaussian.

**Lemma D.3** (Lemma 2.1 of [BS21]). Let $\mathcal{F}$ be any class of functions from $\mathbb{R}^d \rightarrow [-1, 1]$. Let $\{(x_i, y_i)\}_{i=1}^{n}$ be i.i.d. input-output pairs in $\mathbb{R}^d \times [-1, 1]$ for any given norm $||\cdot||$. Assume that the expected conditional variance of the output (i.e., the “noise level”) is strictly positive, denoted by $\sigma^2 := \mathbb{E}[\text{Var}[y|\cdot]] > 0$.

$$\Pr \left( \exists f \in \mathcal{F} : \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \right) \leq 2 \exp \left( -\frac{ne^2}{83} \right) + \Pr \left( \exists f \in \mathcal{F} : \frac{1}{n} \sum_{i=1}^{n} f(x_i)z_i \geq \frac{\epsilon}{4} \right).$$

**Theorem D.4** (Overparametrization is necessary for good data fitting). Let $\mathcal{F}$ be any class of functions from $\mathbb{R}^d \rightarrow [-1, 1]$ with the $J$-Lipschitz parametrization with $w \in \mathcal{W} \subseteq \mathbb{R}^p$. That is, for all parametrized function $f_w$ and $f_{w'}$ in $\mathcal{F}$, $||f_w - f_{w'}||_{\mathcal{F}} \leq J ||w - w'||$, where $w, w' \in \mathcal{W} \subseteq \mathbb{R}^p$ and $\text{diam}(\mathcal{W}) \leq W$. Suppose that $|z_i| \leq 2$ and $f : \mathbb{R}^d \rightarrow [-1, 1]$. Then with high probability, we have

$$\exists f \in \mathcal{F} : \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \Rightarrow p \geq \tilde{\Omega}(ne^2).$$

22
Proof. Since $|f(x_i)z_i| \leq 2$, by Lemma D.1 $f(x_i)z_i$ is 4-subgaussian. Then by Lemma D.2 and the fact that $E[z_i|x_i] = 0$ (thus $E[f(x_i)z_i] = 0$), $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} f(x_i)z_i$ is 72-subgaussian. For any $f$, this yields

$$
\Pr \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f(x_i)z_i \geq t \right) \leq 2 \exp\left( -\frac{(t/c_0)^2}{2} \right),
$$

where $c_0$ is an absolute constant. Choosing $t = \sqrt{n}\epsilon/4$, we have

$$
\Pr \left( \frac{1}{n} \sum_{i=1}^{n} f(x_i)z_i \geq \epsilon \right) \leq 2 \exp\left( -c_1 n\epsilon^2 \right),
$$

where $c_1$ is an absolute constant. Thus by union bound,

$$
\Pr \left( \exists f \in \mathcal{F} : \frac{1}{n} \sum_{i=1}^{n} f(x_i)z_i \geq \epsilon \right) \leq 2 |\mathcal{F}| \exp\left( -c_1 n\epsilon^2 \right) \leq 2 \exp\left( p \ln(60WJ\epsilon^{-1}) - c_1 n\epsilon^2 \right).
$$

This together with Lemma D.3 proves the theorem. \qed

E An Interesting Property on Randomized Smoothing

Because the smoothed classifier in randomized smoothing has an explicit Lipschitz constant, it is interesting to compare its Lipschitz constant with the upper bound derived in [BS21].

**Theorem E.1** (Lemma 1 of [SYL+19]). Denote by $f : \mathbb{R}^d \to [-1, 1]$ the base regression network, in randomized smoothing we create a smoothed encoder $g$ with $f$ and a smoothing distribution $q$, for a sample $x$ we have

$$
g(x) = \mathbb{E}_{z \sim q}[f(x + z)].
$$

If $q \sim \mathcal{N}(0, \sigma^2 I_d)$, $g(x)$ is $\sqrt{\frac{2}{\pi\sigma^2}}$-Lipschitz continuous, i.e., for any $x$ and $y$,

$$
|g(x) - g(y)| \leq \sqrt{\frac{2}{\pi\sigma^2}} \|x - y\|_2.
$$

**Theorem E.2** (Noise level for good data fitting by randomized smoothing). Let $\mathcal{F}$ be a class of base functions from $\mathbb{R}^d \to \mathbb{R}$, and denote by $\mathcal{H} = \{ h : h(x) = \mathbb{E}_{g \sim \mathcal{N}(0, \eta^2 I)} f(x + g) \}$ a class of smoothed functions that are yielded from $\mathcal{F}$. Assume that

- $\mathcal{F}$ admits a parametrization by $p$ real parameters, each of size at most $\text{poly}(n, d)$.
- The distribution $\mu$ of the covariates $x_i$ satisfies isoperimetry.
- The expected conditional variance of the output (i.e., the “noise level”) is strictly positive, denoted $\sigma^2 := \mathbb{E}[\text{Var}[y|x]] > 0$.

Then with high probability over the sampling of the data, one has simultaneously for all $f \in \mathcal{F}$:

$$
\frac{1}{n} \sum_{i=1}^{n} (y_i - h(x_i))^2 \leq \sigma^2 - \epsilon \quad \Rightarrow \quad \eta^2 \leq O\left( \frac{p}{n\sigma^2} \right).
$$

23
Proof. Suppose that $\frac{1}{n} \sum_{i=1}^{n} (y_i - h(x_i))^2 \leq \sigma^2 - \epsilon$. By Theorem 3 in [BS21], we have

$$\text{Lip}(h) \geq \tilde{\Omega}\left(\epsilon \sqrt{\frac{nd}{p}}\right).$$

On the other hand, by Theorem E.1, we have for all $f \in F$:

$$\text{Lip}(h) \leq \sqrt{\frac{2}{\pi \eta^2}}.$$

This yields

$$\eta^2 \leq \tilde{O}\left(\frac{p}{nd \epsilon^2}\right).$$

\[\square\]

F An Inverse Property on Randomized Smoothing

**Theorem F.1.** Let $F$ be a class of base functions from $\mathbb{R}^d \rightarrow \mathbb{R}$, and denote by $\mathcal{H} = \{h : h(x) = \mathbb{E}_{g \sim \mathcal{N}(0, \eta^2 I)} f(x + g)\}$ a class of smoothed functions that are yielded from $F$. Suppose that $\frac{1}{n} \sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \sigma^2 - 3\epsilon$. Assume that the expected conditional variance of the output (i.e., the “noise level”) is strictly positive, denoted $\sigma^2 := \mathbb{E}\{\text{Var}[y|x]\} > 0$. We have

$$\eta \leq O\left(\frac{\epsilon}{d \cdot \text{Lip}(f)}\right) \Rightarrow \frac{1}{n} \sum_{i=1}^{n}(y_i - \mathbb{E}_g f(x_i + g))^2 \leq \sigma^2 - \epsilon.$$

**Proof.**

\[
\frac{1}{n} \sum_{i=1}^{n}(y_i - \mathbb{E}_g f(x_i + g))^2 \\
= \frac{1}{n} \sum_{i=1}^{n}(y_i - f(x_i) + f(x_i) - \mathbb{E}_g f(x_i + g))^2 \\
= \frac{1}{n} \sum_{i=1}^{n}(y_i - f(x_i))^2 + \frac{1}{n} \sum_{i=1}^{n}(f(x_i) - \mathbb{E}_g f(x_i + g))^2 + \frac{2}{n} \sum_{i=1}^{n}(y_i - f(x_i))(f(x_i) - \mathbb{E}_g f(x_i + g)).
\]

As for the first term, we have

$$\frac{1}{n} \sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \sigma^2 - 3\epsilon.$$

As for the second term, we note that

$$f(x_i) - f(x_i + g) \leq \text{Lip}(f) ||g||_2 \Rightarrow f(x_i) - \mathbb{E}_g f(x_i + g) \leq \text{Lip}(f)\mathbb{E}_g ||g||_2 = \text{Lip}(f)\eta \sqrt{2\Gamma((d + 1)/2)} \Gamma(d/2)$$

$$\Rightarrow \frac{1}{n} \sum_{i=1}^{n}(f(x_i) - \mathbb{E}_g f(x_i + g))^2 \leq O(\text{Lip}(f)^2 \eta^2 d),$$

where the last step holds because $\frac{\Gamma((d+1)/2)}{\Gamma(d/2)} \sim d^{1/2}$. As for the third term, note that

$$|y_i - f(x_i)| \leq O(1).$$
Therefore, we have
\[ \frac{2}{n} \sum_{i=1}^{n} (y_i - f(x_i))(f(x_i) - \mathbb{E}_g f(x_i + g)) \leq O(Lip(f)\eta \sqrt{d}). \]

It yields
\[ \eta \leq O \left( \frac{\epsilon}{d \cdot Lip(f)} \right) \Rightarrow \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbb{E}_g f(x_i + g))^2 \leq \sigma^2 - \epsilon. \]