A Law of Robustness beyond Isoperimetry

Yihan Wu [1] **Heng Huang** [1] **Hongyang Zhang** [2]

Abstract

We study the *robust interpolation problem* of arbitrary data distributions supported on a bounded space and propose a two-fold law of robustness. Robust interpolation refers to the problem of interpolating n noisy training data points in $\mathbb{R}^d$ by a Lipschitz function. Although this problem has been well understood when the samples are drawn from an isoperimetry distribution, much remains unknown concerning its performance under generic or even the worst-case distributions. We prove a Lipschitzness lower bound $\Omega(\sqrt{n/p})$ of the interpolating neural network with p parameters on arbitrary data distributions. With this result, we validate the law of robustness conjecture in prior work by Bubeck, Li, and Nagaraj on two-layer neural networks with polynomial weights. We then extend our result to arbitrary interpolating approximators and prove a Lipschitzness lower bound $\Omega(n^{1/d})$ for robust interpolation. Our results demonstrate a two-fold law of robustness: i) we show the potential benefit of overparametrization for smooth data interpolation when $n = \mathrm{poly}(d)$, and ii) we disprove the potential existence of an $\mathcal{O}(1)$-Lipschitz robust interpolating function when $n = \exp(\omega(d))$.

1. Introduction

Robustness has been a central research topic in machine learning (Szegedy et al., 2014; Goodfellow et al., 2014), statistics (Huber, 2004), operation research (Ben-Tal et al., 2009), 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* (Bubeck & Sellke, 2023). That is, given noisy training data $\{(x_i, g(x_i) + z_i)\}_{i=1}^n$ of size n where $x_1, \cdots, x_n$ are restricted in a unit ball and $z_1, \cdots, z_n$ have variance > 0, how many network parameters and training samples are needed for robust interpolation 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 (Bubeck & Sellke, 2023): 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 (Northcutt et al., 2021). 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 (Bubeck & Sellke, 2023; Bubeck et al., 2021), many fundamental questions remain unresolved. In modern learning theory, it was commonly believed that 1) big data (Schmidt et al., 2018), 2) low dimensionality of input (Blum et al., 2020; Yang et al., 2020a; Kumar et al., 2020), and 3) overparametrization (Bubeck & Sellke, 2023; Bubeck et al., 2021) 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 empirical evidence to indicate that enlarging the model size (overparametrization) improves robustness when n is moderately large (*e.g.*, when $n = \mathrm{poly}(d)$, see (Madry et al., 2017; Schmidt et al., 2018)). Our work verifies the benefit of overparametrization for fitting a neural network with p parameters below the noise level by proving such neural networks must have a Lipschitzness lower bound $\Omega(\sqrt{n/p})$. On the other hand, big

[1]Department of Computer Science, University of Maryland at College Park [2]School of Computer Science, University of Waterloo. Correspondence to: Yihan Wu <ywu42@umd.edu>, Heng Huang <heng@umd.edu>, Hongyang Zhang <hongyang.zhang@uwaterloo.ca>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

data and large models may not be a remedy for robustness if n goes even larger. We show that for any approximator, no matter how many parameters it contains, its Lipschitzness is of order $\Omega(n^{1/d})$. In particular, our result disproves the existence of learning an $\mathcal{O}(1)$-Lipschitz function with $n = \exp(\omega(d))$. Besides, by showing that for any learning algorithm, there exists a joint data distribution such that one needs at least $n = \exp(\Omega(d))$ samples to learn an $\mathcal{O}(1)$-Lipschitz function with good population error, we demonstrate that big data are also necessary for robust interpolation in some special cases.

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 (Bubeck & Sellke, 2023). A probability measure μ on $\mathbb{R}^d$ satisfies c-isoperimetry if for any bounded L-Lipschitz $f : \mathbb{R}^d \to \mathbb{R}$, and any $t \geq 0$,

$$\Pr(|f(x) - \mathbb{E}[f(x)]| \geq t) \leq 2\exp(-\frac{dt^2}{2cL^2}).$$

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 $\mathcal{N}(0, \frac{I_d}{d})$, 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 3.4 go beyond isoperimetry and providing a lower bound of robustness for functions with p parameters under *arbitrary* distributions in the bounded space. Our results of Theorem 3.9 go even further by providing a universal lower bound of robustness for *any* model class, including the class of neural networks with arbitrary architecture.

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, in our work we use the mean squared error as in Bubeck & Sellke (2023), i.e., $l(f(x), y) = (f(x) - y)^2$. Let $\mathcal{L}_{\mathcal{D}}(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(f(x), y)]$ 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 $\|x\|$ a generic norm. Examples of norms include $\|x\|_\infty$, the infinity norm, and $\|x\|_2$, the ℓ_2 norm. We will frequently use $(\mathcal{X}, \|\cdot\|)$ to represent the normed linear space of $\mathcal{X}$ with norm $\|\cdot\|$. Define $\mathrm{diam}(\mathcal{X})$ as the diameter of $\mathcal{X}$ w.r.t. the norm $\|\cdot\|$. For a given score function f, we denote by $\mathrm{Lip}_{\|\cdot\|}(f)$ (or sometimes $\mathrm{Lip}(f)$ 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.

1.1. Our results

Our law of robustness is two-fold: a) overparametrization can potentially help robust interpolation when $n = \mathrm{poly}(d)$ (Section 3.1), and b) there exists no robust interpolation when $n = \exp(\omega(d))$ (Section 3.2).

Lipschitzness (or local Lipschitzness) is an important characterization of adversarial robustness for learning algorithms (Yang et al., 2020b; Zhang et al., 2019; Wu et al., 2022b;c). The popular randomized smoothing approaches (Cohen et al., 2019; Li et al., 2019; Wu et al., 2022d) can provide robust guarantee through Lipschitzness but suffer curse of dimensionality problem (Wu et al., 2021). Thus, studying the Lipschitzness is crucial for understanding robustness. For a given score function f, we denote by $\mathrm{Lip}_{\|\cdot\|}(f)$ the Lipschitz constant of f *w.r.t.* the norm $\|\cdot\|$. That is, for any x_1, x_2 in the input space, $|f(x_1) - f(x_2)| \leq \mathrm{Lip}_{\|\cdot\|}(f)\|x_1 - x_2\|$. Our results show lower bounds 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 ℓ_∞ robust radius of 0.031 or 0.062 (Zhang et al., 2019; Madry et al., 2017), which corresponds to a (local) Lipschitz constant of $\mathcal{O}(1)$ for the classifier.

Overparametrization may benefit robust interpolation. The universal law of robustness by Bubeck & Sellke (2023) provides an $\Omega(\sqrt{nd/p})$ Lipschitzness lower bound of the interpolating functions when the underlying distribution is isoperimetry (see Theorem 2.1). Our first result goes beyond the isoperimetry assumption, and provides an $\Omega(\sqrt{n/p})$ Lipschitzness lower bound of the interpolating functions under arbitrary distribution. We note that the $\sqrt{d}$ difference between the two Lipschitzness lower bounds is due to the special property of the isoperimetry assumption (see Remark 3.5). Our result predicts the *potential* existence of an $\mathcal{O}(1)$-Lipschitz function that fits the data below the noise level when $p = \Omega(n)$. The following informal theorem illustrates the results (the detailed theorems are introduced at later sections):

Theorem A (informal version of Theorem 3.4). *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 $\{x : \|x\| \le 1\} \times [-1,1]$ for any given norm $\|\cdot\|$. Assume that:

1. *The expected conditional variance of the output (i.e., the "noise level") is strictly positive, denoted by $\sigma^2 := \mathbb{E}[\mathrm{Var}[y|x]] > 0$.*
2. *$\mathcal{F}$ admits a J-Lipschitz parametrization by p real parameters, each of size at most $poly(n,d)$.*

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

$$\frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \le \sigma^2 - \epsilon \Rightarrow \mathrm{Lip}_{\|\cdot\|}(f) \ge \Omega\left(\epsilon\sqrt{\frac{n}{p}}\right).$$

Remark 1.1. Our theorem takes a further step in proving the Conjecture 1 in Bubeck et al. (2021), where it is conjectured that for generic data sets, with high probability, any f in the collections of two layer networks with p parameters fitting the data must also satisfy $\mathrm{Lip}_{\|\cdot\|}(f) \ge \Omega(\sqrt{n/p})$. We validate the conjecture under the polynomial weights assumption, where Bubeck & Sellke (2023) validate the Conjecture 1 under the polynomial weights assumption and the isoperimetry assumption.

Remark 1.2 (Strong overparametrization is not necessary for the robust interpolation). The Lipschitzness lower bound of Bubeck & Sellke (2023) suggests strong overparametrization, i.e., $p = \Omega(nd)$, is required for the robust interpolation under the isoperimetry assumption. Our theorem shows that strong overparametrization may not be a necessary condition for the robust interpolation on a general distribution. Moderate overparametrization with $p = \Omega(n)$ may also be enough for robust interpolation. Our results are consistent with the empirical observations that CIFAR10 (50000 images) can be robustly fitted by a model with $p = 10^6$, and ImageNet (10^7 images) can be robustly fitted by a model with $p = 10^7 \sim 10^8$.

Big data hurts robust interpolation. Under the assumptions of isoperimetry distribution and the J-Lipschitz parameterized functions, the universal law of robustness by Bubeck & Sellke (2023) predicts the *potential* existence of an $\mathcal{O}(1)$-Lipschitz function fits the data below the noise level when $p = \Omega(nd)$. Our result goes beyond the two assumptions and disproves the existence of such $\mathcal{O}(1)$-Lipschitz functions in the big data scenario when $n = \exp(\omega(d))$ for *arbitrary* distributions:

Theorem B (informal version of Theorem 3.9). *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 $\{x : \|x\| \le 1\} \times [-1,1]$ for any given norm $\|\cdot\|$. Assume that:*

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

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

$$\frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \le \sigma^2 - \epsilon \;\Rightarrow\; \mathrm{Lip}_{\|\cdot\|}(f) \ge \Omega(\epsilon n^{1/d}).$$

Difference between our results and Bubeck & Sellke (2023). Bubeck & Sellke (2023) proposed a universal law of robustness for general class of functions (see Theorem 2.1). Our results Theorem 3.4 and Theorem 3.9 share the same setting with Theorem 2.1, while the former ones make much weaker assumptions: 1) Both Theorem 3.4 and Theorem 3.9 do not require an isoperimetry assumption of input distributions. 2) Theorem 3.9 does not make any assumption on the Lipschitzness and size of model parametrization. Moreover, while Theorem 2.1 predicts potential existence of an $\mathcal{O}(1)$-Lipschitz robust interpolating function when $p = \Omega(nd)$, Theorem 3.9 disproves the hypothesis in the big data scenario when $n = \exp(\omega(d))$ for *arbitrary* distributions in the bounded space. Besides, our bounds work for all $\ell_p (p \ge 1)$ norm while the bound in Bubeck & Sellke (2023) only focuses on ℓ_2 norm.

Practical implications. Our analysis provides important implications for practical settings. When selecting the models for learning on a certain dataset, ideally the number of parameters in the selected model should be the same (or slightly larger) scale of the dataset in order to get good robust performance. When the size of dataset is too large comparing to the dimension of dataset, in order to achieve good robustness, it may be beneficial to either reduce the size of the training data or scatter the data in a higher-dimensional space by padding special covariates. This approach can help to mitigate the negative effects of the curse of big data and improve model robustness, particularly when dealing with large datasets in practical applications (Wu et al., 2022a; 2023; Sun et al., 2021; 2022).

2. Related Work

Robust interpolation problem. Bubeck et al. (2021) provided the first guarantee on the law of robustness for two-layer neural networks which was later extended by Bubeck & Sellke (2023) to a universal law of robustness for general class of functions under isoperimetry distributions. A probability measure μ on $\mathbb{R}^d$ satisfies c-isoperimetry if for any bounded L-Lipschitz $f : \mathbb{R}^d \to \mathbb{R}$, and any $t \ge 0$, $\Pr(|f(x) - \mathbb{E}[f(x)]| \ge t) \le 2\exp(-\frac{dt^2}{2cL^2})$.

Theorem 2.1 (Theorem 1 of Bubeck & Sellke (2023)). *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. *The expected conditional variance of the output (i.e., the "noise level") is strictly positive, denoted by $\sigma^2 := \mathbb{E}[\mathrm{Var}[y|x]] > 0$.*

2. *$\mathcal{F}$ admits a J-Lipschitz parametrization by p real parameters, each of size at most $poly(n, d)$.*
3. *The distribution μ of the input x_i satisfies isoperimetry (or a mixture thereof).*

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

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \Rightarrow \mathrm{Lip}_{\|\cdot\|_2}(f) \geq \Omega\left(\epsilon\sqrt{\frac{nd}{p}}\right).$$

Our work extends the result of Bubeck & Sellke (2023) by consequently removing the third assumption (see Theorem 3.4) and the second assumption (see Theorem 3.9).

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, Bhattacharjee et al. (2021) considered the sample complexity of robust linear classification on the separated data. Yin et al. (2019) studied the adversarially robust generalization problem through the lens of Rademacher complexity. Cullina et al. (2018) extended the PAC-learning framework to account for the presence of an adversary. Montasser et al. (2019) 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. Schmidt et al. (2018) showed an $\Omega(\sqrt{d})$-factor gap between the standard and robust sample complexity for a mixture of Gaussian distributions in ℓ_∞ robustness, which was later extended to the case of ℓ_p robustness with a tight bound by Bhagoji et al. (2019); Dobriban et al. (2020); Dan et al. (2020). 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.

3. A Two-fold Law of Robustness

In this section, we present our main theoretical analysis, which contributes to our two-fold law of robustness. All missing proofs can be found in the appendix.

Robust interpolation problem. We first introduce our problem settings. Given noisy training data $\{(x_i, y_i := g(x_i) + z_i)\}_{i=1}^n$ of size n where $x_1, \ldots, x_n$ are training samples, $g(x_1), \ldots, g(x_n)$ the ground truth, and $z_1, \ldots, z_n$ have variance $\sigma^2 > 0$, we say a model f robustly interpolates (or fits the data below the noise level) the training data if and only if

$$\exists \epsilon > 0, \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \sigma^2 - \epsilon.$$

Our two-fold law of robustness. a) Overparametrization can potentially help robust interpolation when $n = \mathrm{poly}(d)$ (Section 3.1); b) There exists no robust interpolation when $n = \exp(\omega(d))$ (Section 3.2).

3.1. A Lipschitz lower bound beyond the isoperimetry assumption.

In this part, we show the first part of our two-fold law of robustness: overparametrization can potentially help robust interpolation when $n = \mathrm{poly}(d)$. Notice, here we claim "potentially help" as overparametrization is only a necessary but not sufficient condition for robust interpolation.

Motivation. We notice that, the proof of Theorem 2.1 (Bubeck & Sellke, 2023) depends heavily on the definition of isoperimetry distribution, i.e., $\Pr(|f(x) - \mathbb{E}[f(x)]| \geq t) \leq 2\exp(-\frac{dt^2}{2cL^2})$ for L-Lipschitz $f : \mathbb{R}^d \to \mathbb{R}$. This formula indicates the high-concentration property of isoperimetry distributions due to the $\exp(-d)$ dependency of $\Pr(|f(x) - \mathbb{E}[f(x)]| \geq t)$. The $\exp(-d)$ dependency is also the reason that the Lipschitzness lower bound of Bubeck & Sellke (2023) is $\Omega(\sqrt{nd/p})$ instead of the $\Omega(\sqrt{n/p})$ lower bound we derived.

Challenge. One may naturally come up with the idea to derive a bound of $\Pr(|f(x) - \mathbb{E}[f(x)]| \geq t)$ for arbitrary distributions and go beyond the isoperimetry distribution. However, the challenge is that unlike the regular concentration bound on $\Pr(|x - \mathbb{E}[x]| \geq t)$, we are dealing with a more complicate case, where the random variable is $f(x)$ with arbitrary L-Lipschitz f. To solve this problem, we apply the Azuma's inequality below:

Lemma 3.1 (Azuma's inequality (Azuma, 1967))**.** *Suppose $\{X_k : k = 0, 1, 2, 3, \ldots\}$ is a martingale and $|X_k - X_{k-1}| \leq c_k$ almost surely. Then for all positive integers N and $\epsilon > 0$,*

$$\Pr(|X_N - X_0| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2\sum_{k=1}^{N} c_k^2}\right).$$

Azuma's inequality shows the concentration bound for the values of martingales that have bounded differences. With this lemma, we are able to derive the following concentration bound for arbitrary distributions on a bounded space.

Lemma 3.2. *Given an arbitrary probability measure μ on the bounded space $\mathcal{X} \subset \mathbb{R}^d$, for any L-Lipschitz $f : \mathbb{R}^d \to \mathbb{R}$, and any $t \geq 0$,*

$$\Pr(|f(x) - \mathbb{E}[f(x)]| \geq t) \leq 2\exp(-\frac{t^2}{2\mathrm{diam}(\mathcal{X})^2 L^2}).$$

Comparing with the $\exp(-\frac{dt^2}{2cL^2})$ bound for the isoperimetry distributions, our bound for arbitrary distributions only differs a d on the numerator of the term inside the exponential. In order to achieve the same concentration bound of isoperimetry distributions, one need $\mathrm{diam}(\mathcal{X}) = \Theta(1/\sqrt{d})$,

which means our input are located on an $\Theta(1/\sqrt{d})$-diameter space. As the real world datasets are usually supported on an $\Theta(1)$-diameter space, matching the isoperimetry bound for all distributions is empirical meaningless.

With Lemma 3.2, we can start to calculate the Lipschitzness lower bound with the following lemma on finite function class

Lemma 3.3. *Let $\mathcal{F}$ be a finite class of L-Lipschitz 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 $\{x : \|x\| \le 1\} \times [-1,1]$ for any given norm $\|\cdot\|$. Assume that the expected conditional variance of the output (i.e., the "noise level") is strictly positive, denoted by $\sigma^2 := \mathbb{E}[\mathrm{Var}[y|x]] > 0$, we have*

$$\Pr\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \le \sigma^2 - \epsilon\right) \le 4\exp\left(-\frac{n\epsilon^2}{8^3}\right) + |\mathcal{F}|\exp\left(-\frac{\epsilon^2 n}{2^{10}L^2}\right).$$

Lemma 3.3 shows the connection between the robust interpolation problem and the Lipschitzness of the underlying functions. Notice, the probability of $\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \le \sigma^2 - \epsilon$ decreases with L, which indicates that we need a large enough L to make sure that there exists f satisfying the condition of robust interpolation problem. With this intuition, we can calculate the following Lipschitzness lower bound for the robust interpolation problem without the isoperimetry assumption.

Theorem 3.4. *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 $\{x : \|x\| \le 1\} \times [-1,1]$ for any given norm $\|\cdot\|$. Assume that:*

1. *The expected conditional variance of the output (i.e., the "noise level") is strictly positive, denoted by $\sigma^2 := \mathbb{E}[\mathrm{Var}[y|x]] > 0$.*
2. *J-Lipschitz parametrization: $\mathcal{F} = \{f_w, w \in \mathcal{W}\}$ with $\mathcal{W} \subset \mathbb{R}^p$, $\mathrm{diam}(\mathcal{W}) \le W$ and for any $w_1, w_2 \in W$,*

$$||f_{w_1} - f_{w_2}||_{\mathcal{F}} \le J||w_1 - w_2||.$$

Then, with probability at least $1-\delta$, one has simultaneously for all $f \in \mathcal{F}$:

$$\frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \le \sigma^2 - \epsilon \Rightarrow \mathrm{Lip}_{\|\cdot\|}(f) \ge \frac{\epsilon}{32}\sqrt{\frac{n}{p\ln(36WJ\epsilon^{-1}) + \ln(2/\delta)}}.$$

The crucial part of the proof is to find a finite $\epsilon/6J$-covering of $\mathcal{F}$ with the J-Lipschitz parametrization assumption. Then we can apply Lemma 3.3 to this finite covering set and get the Lipschitzness lower bound. We will show in the next section that without the J-Lipschitz parametrization assumption, one can hardly use the similar proof technique to derive the Lipschitzness lower bound.

Bubeck & Sellke (2023) showed that under neural network settings, J is always of polynomial order of the diameter of the weight space. Thus, if the weight is only polynomial large w.r.t. d and n, $\ln(60WJ\epsilon^{-1})$ would not affect the Lipschitzness bound too much and we may neglect it in its asymptotic approximation. Thus, we have a Lipschitzness lower bound of order $\Omega(\epsilon\sqrt{n/p})$ for the robust interpolation problem. Our theorem validates the first part of our law of robustness, i.e., the potential existence of robust interpolating functions under the overparametrization scenario when $n = \mathrm{poly}(d)$ (see Remark 3.10).

Tightness of our bound. When $n = \mathrm{poly}(d)$, Theorem 4 of Bubeck et al. (2021) has already demonstrated the existence of an at most $\mathcal{O}(\sqrt{n/p})$-Lipschitz two layer network, which fits generic data below the noise level. Thus, our Lipschitzness lower bound is tight.

Remark 3.5 (Difference of the $\sqrt{d}$-dependency between Theorem 2.1 and 3.4). Comparing to the $\Omega(\epsilon\sqrt{nd/p})$ of Lipschitzness lower bound in Theorem 2.1, our bound does not depend on the dimension d. This difference, as we discussed in Lemma 3.2, is due to the isoperimetry assumption. In Bubeck et al. (2021), it's also showed that the tight Lipschitzness lower bound of two layer networks is of order $\Omega(\epsilon\sqrt{n/p})$, which is consistent with our results.

3.2. A Lipschitz lower bound beyond the J-Lipschitz parametrization assumption.

In this part, we show the second part of our two-fold law of robustness. We demonstrate an intriguing observation that huge data hurts robust interpolation. Our analysis leads to a universal lower bound of Lipschitzness regarding the robust interpolation problem, which goes beyond the isoperimetry and J-Lipschitz parametrization assumptions. Our analysis is based on the relation between Rademacher complexity and the generalization gap between the population error $\mathcal{L}_{\mathcal{D}}(f)$ and the training error $\mathcal{L}_S(f)$.

Motivation. The J-Lipschitzness parametrization assumption provides us a simple way to find a covering of the function space $\mathcal{F}$. Although the Lipschitzness lower bound in Bubeck & Sellke (2023) has only logarithmic dependency with respect to J, it may still affect the Lipschitzness lower bound when the weight of neural networks is exponentially large w.r.t. d, or the number of layers of neural works is polynomial w.r.t. d. Thus, we seek to derive a Lipschitzness lower bound beyond the J-Lipschitzness parametrization assumption.

Challenge. Without the J-Lipschitzness parametrization

assumption, the covering number of the function $\mathcal{F}$ will have more complicate dependency on the Lipschitzness L (see Lemma 3.6). In this case, calculating Lipschitzness lower bound with Lemma 3.6 and Lemma 3.3 requires one to solve an inequality like $L^{-d}\ln L + L^{-2} \geq C$, which obviously has no closed-form solution when $d \geq 3$. Thus, we need other techniques to deal with this case. Recall the objective of robust interpolation problem is $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \sigma^2 - \epsilon$, one can immediately find that the left hand side formula is the train error with mean squared loss $\mathcal{L}_S(f)$. Under the label noise settings, we have $\mathcal{L}_{\mathcal{D}}(f) = \mathbb{E}_{\mathcal{D}}[(f(x)-y)^2] \geq \mathbb{E}_x[\mathrm{Var}(y|x)] = \sigma^2$, which yields $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \Rightarrow \mathcal{L}_S(f) \leq \mathcal{L}_{\mathcal{D}}(f) - \epsilon$. Therefore, if one can derive

$$\mathcal{L}_S(f) \leq \mathcal{L}_{\mathcal{D}}(f) - \epsilon \Rightarrow \mathrm{Lip}_{\|\cdot\|}(f) \geq \Omega(\epsilon n^{1/d}),$$

a natural corollary is that

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \Rightarrow \mathcal{L}_S(f) \leq \mathcal{L}_{\mathcal{D}}(f) - \epsilon \Rightarrow \mathrm{Lip}_{\|\cdot\|}(f) \geq \Omega(\epsilon n^{1/d}).$$

In this way, we successfully convert the robust interpolation problem to a generalization problem between the empirical error and population error under the mean squared loss, which can be solved by the statistical learning techniques, e.g., VC dimension and Rademacher complexity. We focus on the Rademacher complexity in this part.

Rademacher complexity. We start with the definition of Rademacher complexity, which measures the richness of a function class. For a set $\mathcal{A} \subset \mathbb{R}^n$, the Rademacher complexity is defined as

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

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

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

For every function $f \in \mathcal{F}$, the generation error between $\mathcal{L}_{\mathcal{D}}(f)$ 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 $f\in\mathcal{F}$,

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

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

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

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 Shalev-Shwartz & Ben-David (2014)) 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 (von Luxburg & Bousquet, 2004) that the Rademacher complexity of a set is directly related to the number of ϵ-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 η, let $N(\mathcal{X},\eta,||\cdot||)$, a.k.a. the η-covering number, be the minimum number of η-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 upper bound of the covering number of $\mathcal{F}$:

Lemma 3.6 (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 $\mathrm{Lip}_{||\cdot||}(f) \leq L$. If $\mathcal{X}$ is connected and centered, we have for every $\epsilon > 0$,*

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

The Dudley's integral provides the relation between the covering number of a function class and its Rademacher complexity. With Dudley's integral, von Luxburg & Bousquet (2004) showed that for every $\epsilon > 0$,

$$\mathbb{E}_{S'\in\mathcal{D}^n}[R(B_L\circ S)] \leq 2\epsilon + \frac{4\sqrt{2}}{\sqrt{n}}\int_{\epsilon/4}^{\mathrm{diam}(B_L)}\sqrt{\ln(N(B_L,u,||\cdot||_{\mathcal{F}}))}du. \tag{3}$$

Notice that when $u > 2L\cdot\mathrm{diam}(\mathcal{X})$, the number of u-covering is 1 and $\ln(N(B_L,u,||\cdot||_{\mathcal{F}})) = 0$. Combining it with Lemma 3.6 yields the following lemma:

Lemma 3.7. *Let $(\mathcal{X}, ||\cdot||)$ be a bounded and connected space and B_L be all functions $f \in \mathcal{F}$ with $\mathrm{Lip}_{||\cdot||}(f) \le L$. Let $n = |S|$. If $\mathcal{X}$ is connected and centered, for any $\epsilon > 0$*

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

As all the variables in Lemma 3.7 are known, by calculating the integration, one can derive an upper bound of Rademacher complexity $\mathbb{E}_{S\in\mathcal{D}^n}[R(B_L \circ S)]$:

Lemma 3.8. *If $\mathrm{diam}(\mathcal{X}) = 2$ w.r.t. $||\cdot||$ and $d \ge 3$, we have*

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

According to Equation 1 in (Mendelson & Vershynin, 2003), when $\frac{u}{2L} \le \mathrm{diam}(\mathcal{X})$, $N(\mathcal{X}, \frac{u}{2L}, ||\cdot||) \le (\frac{6L\cdot\mathrm{diam}(\mathcal{X})}{u})^d$ if $\mathcal{X} \subseteq \mathbb{R}^d$. Then the integral part will be $\sqrt{(\frac{12L}{u})^d \ln 2 + \ln\lceil \frac{2L\cdot\mathrm{diam}(\mathcal{X})}{u} \rceil}$, which is no more than $\sqrt{(\frac{12L}{u})^d \ln 2} + \sqrt{\ln(\frac{4L}{u}+1)}$. Taking $\epsilon = \Theta(\frac{L}{n^{1/d}})$, the integral part will be bounded by $\Theta(Ln^{1/2-1/d})$. Thus $\mathbb{E}_{S'\in\mathcal{D}^n}[R(B_L \circ S')] \le \Theta(\frac{L}{n^{1/d}}) + \frac{4\sqrt{2}}{\sqrt{n}}\Theta(Ln^{1/2-1/d}) = \Theta(\frac{L}{n^{1/d}})$.

In our settings, we are interested in the squared ℓ_2 loss $l(f(x), y) = (f(x) - y)^2$. We have $\nabla_{f(x)} l(f(x), y) = 2(f(x) - y) \le 2(|f(x)| + |y|) \le 4$, i.e., $l(f(x), y)$ is 4-Lipschitz w.r.t. $f(x)$ for arbitrary $y \in [-1, 1]$. Thus, $\mathbb{E}_{S\in\mathcal{D}^n}[R(l \circ B_L \circ S)] \le 4\mathbb{E}_{S\in\mathcal{D}^n}[R(B_L \circ S)] = \mathcal{O}\left(\frac{L}{n^{1/d}}\right)$. Combining this result with Equation 1 yields the main theorem of our paper:

Theorem 3.9 (Lipschitzness Lower Bound Beyond the J-Lipschitz parametrization assumption)**.** *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 $\{x : \|x\| \le 1\} \times [-1, 1]$ for any given norm $\|\cdot\|$. Assume that:*

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

Then with probability at least $1 - \delta$, for all $f \in \mathcal{F}$:

$$\frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \le \sigma^2 - \epsilon \Rightarrow \mathrm{Lip}_{\|\cdot\|}(f) \ge \frac{n^{1/d}}{K}\left(\frac{1}{8}\epsilon - \frac{1}{2}\sqrt{\frac{2\ln(2/\delta)}{n}}\right),$$

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

Theorem 3.9 states that, for all data distribution $\mathcal{D}$ with label noise of variance σ^2 and every function $f : \mathcal{X} \to [-1, 1]$, overfitting i.e. $\frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \le \sigma^2 - \epsilon$ implies $\mathrm{Lip}_{\|\cdot\|} \ge \Omega(\epsilon n^{1/d})$, which validates the second part of our law of robustness, i.e., achieving good robust interpolation is impossible when $n = \exp(\omega(d))$.

Remark 3.10. Theorem 3.9 disprove the existence of robust interpolating functions when $n = \exp(\omega(d))$. Thus, the first part of our law of robustness holds only when $n = \mathrm{poly}(d)$.

Tightness of our bound. 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 evenly in the space, where the distance between any training pair is as large as $\Theta(1/n^{1/d})$. Inspired by this, we complement Theorem 3.9 with a matching Lipschitzness upper bound of $\mathcal{O}(n^{1/d})$, which shows that the Lipschitzness lower bound in Theorem 3.9 is achievable by a certain function and training data:

Theorem 3.11 (Tightness of our bound)**.** *For any distribution $\mathcal{D}$ which is supported on $\{x \in \mathbb{R}^d : ||x|| \le 1\}$, there exist n training samples $\{x_1, ..., x_n\}$ such that $\forall i, j, i \ne j, ||x_i - x_j|| \ge \frac{1}{n^{1/d}}$. 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 3.11 shows that there exists n samples, such that the function which perfectly fits the training samples is $\mathcal{O}(n^{1/d})$-Lipschitz.

3.2.1. Our (counter-intuitive) implications

It was widely believed that 1) big data (Schmidt et al., 2018), 2) low dimensionality of input (Blum et al., 2020), and 3) overparametrization (Bubeck & Sellke, 2023; Bubeck et al., 2021; Gao et al., 2019) improve robustness. Our main results of Theorem 3.9 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 3.9 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 ℓ_∞ 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 3.9 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.

4. Small Data May Hurt Performance and Robustness

In Section 3, we mainly focus on the robust interpolation problem on the training samples. The lower bound given by Theorem 3.9 implies that one can sample at most $\exp(\mathcal{O}(d))$ training samples in order to obtain an $\mathcal{O}(1)$-Lipschitz function in the robust interpolation problem. In this section, we show that $n = \exp(\Omega(d))$ is a necessary condition for obtaining a good population error by any $\mathcal{O}(1)$-Lipschitz learning algorithm.

We now provide a complementary result of Section 3.2. 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 4.1. *Let $\mathcal{A}(S) : \mathcal{X} \to \{-a, a\}$ be any learning algorithm with respect to the squared ℓ_2 loss over a domain $\mathcal{X}$ and samples S. Assume there are label noise $\mathbb{E}[\mathrm{Var}[y|x]] = \sigma^2$. Let m be any number smaller than $|\mathcal{X}|/2$, representing the size of a training set. Then, for any $a > 0$ there exists a distribution $\mathcal{D}$ (with label noise) over $\mathcal{X} \times \{-a, a\}$ such that*

$$\mathbb{E}_{S\sim\mathcal{D}^m}[\mathcal{L}_\mathcal{D}(\mathcal{A}(S))]] \geq \frac{1}{2}(a^2 + \sigma^2).$$

In the next lemma, 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 $\mathcal{D}$. On one hand, if the minimum distance is less than ϵ, we can assign the two samples that achieve the minimum distance with labels 1 and -1, respectively. As the algorithm $\mathcal{A}$ is L-Lipschitz, the maximum difference between the predicted labels of the two selected points is $L\epsilon$. Thus, the error of $\mathcal{A}$ will be larger than $1 - L\epsilon$. On the other hand, if the minimum distance is larger than ϵ, the maximum number of points in the distribution $\mathcal{D}$ will be less than the number of the ϵ-packing of the input space $\mathcal{X}$. By Lemma 4.1, there exists a distribution such than if the number of training samples is less than half of the ϵ-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:

Lemma 4.2 (No-free-lunch theory with L-Lipschitz algorithms)**.** *Let $\mathcal{A}(S) : \mathcal{X} \to [-1, 1]$ 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 ℓ_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 $\sigma^2 := \mathbb{E}_\mathcal{D}[\mathrm{Var}(y|x)] \leq 1/2$. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times [-1, 1]$ with noisy labels such that for all L-Lipschitz (w.r.t. norm $\|\cdot\|$) learning algorithm and any $\epsilon \in [0, \frac{1}{2L}]$:*

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

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

Now we are ready to prove our main theorem.

Theorem 4.3. *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 $\mathcal{L}_\mathcal{D}(f) := \mathbb{E}_\mathcal{D}[(f(x) - y)^2]$ the squared ℓ_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 $\sigma^2 := \mathbb{E}[\mathrm{Var}[y|x]]$. Let $\mathcal{A}(S) : \mathcal{X} \to \mathbb{R}$ be any L-Lipschitz learning algorithm over a training set S. Then there exists a distribution $\mathcal{D}'$ of (x, y) such that*

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

Proof. Consider $\mathcal{X} = \{x \in \mathbb{R}^d : ||x|| \leq 1\}$. We have $M(\mathcal{X}, \eta, ||\cdot||) \geq \left(\frac{1}{\eta}\right)^d$. Thus by Lemma 4.2, there exists a distribution $\mathcal{D}$ such that if $\sigma^2 \leq 0.5$,

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

Taking $\eta = \frac{1/2-\epsilon}{L}$ where $\epsilon \in (0, 1/2)$, we have $n < \frac{1}{2}\left(\frac{2L}{1-2\epsilon}\right)^d$ implies $\mathbb{E}_{S\sim\mathcal{D}^n}[\mathcal{L}_\mathcal{D}(\mathcal{A}(S))] \geq \min\left\{\frac{1}{4}, \epsilon\right\} + \sigma^2$. Thus in the worst case, n has to be at least $\exp(\Omega(d))$ if one wants to achieve good astuteness by any learning algorithm that returns an $\mathcal{O}(1)$-Lipschitz function. This completes the proof of Theorem 4.3. □

Theorem 4.3 states that for certain distributions, n has to be at least $\exp(\Omega(d))$ if one wants to achieve good population error by any $\mathcal{O}(1)$-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 3.9.

5. Conclusions

In this work, we study the robust interpolation problem beyond the isoperimetry assumption, and propose a two-fold law of robustness. We show the potential benefit of overparametrization for smooth data interpolation when $n = \mathrm{poly}(d)$, and disprove the potential existence of an $\mathcal{O}(1)$-Lipschitz robust interpolating function when $n = \exp(\omega(d))$. Besides, we also prove that small data ($\exp(\mathcal{O}(d))$) may hurt robustness on certain distributions. Perhaps surprisingly, the results shed light on the curse of big data and the blessing of dimensionality regarding robustness.

Acknowledgement

Hongyang Zhang is supported by NSERC Discovery Grant RGPIN-2022-03215, DGECR-2022-00357. Yihan Wu and Heng Huang were partially supported by NSF IIS 1838627, 1837956, 1956002, 2211492, CNS 2213701, CCF 2217003, DBI 2225775.

References

Azuma, K. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. *Robust optimization*. Princeton university press, 2009.

Bhagoji, A. N., Cullina, D., and Mittal, P. Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems*, 2019.

Bhattacharjee, R., Jha, S., and Chaudhuri, K. Sample complexity of robust linear classification on separated data. In *International Conference on Machine Learning*, pp. 884–893, 2021.

Blum, A., Dick, T., Manoj, N., and Zhang, H. Random smoothing might be unable to certify ℓ_∞ robustness for high-dimensional images. *Journal of Machine Learning Research*, 21:1–21, 2020.

Bubeck, S. and Sellke, M. A universal law of robustness via isoperimetry. *Journal of the ACM*, 70(2):1–18, 2023.

Bubeck, S., Li, Y., and Nagaraj, D. M. A law of robustness for two-layers neural networks. In *Annual Conference on Learning Theory*, volume 134, pp. 804–820, 2021.

Case, B. M., Gallagher, C., and Gao, S. A note on sub-gaussian random variables. *Cryptology ePrint Archive*, 2019.

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. Certified adversarial robustness via randomized smoothing. *ICML*, 2019.

Cullina, D., Bhagoji, A. N., and Mittal, P. PAC-learning in the presence of evasion adversaries. In *Advances in Neural Information Processing Systems*, pp. 230–241, 2018.

Dan, C., Wei, Y., and Ravikumar, P. Sharp statistical guarantees for adversarially robust gaussian classification. In *International Conference on Machine Learning*, pp. 2345–2355, 2020.

Dobriban, E., Hassani, H., Hong, D., and Robey, A. Provable tradeoffs in adversarially robust classification. *arXiv preprint arXiv:2006.05161*, 2020.

Gao, R., Cai, T., Li, H., Hsieh, C.-J., Wang, L., and Lee, J. D. Convergence of adversarial training in overparametrized neural networks. *Advances in Neural Information Processing Systems*, 32:13029–13040, 2019.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2014.

Huber, P. J. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

Kumar, A., Levine, A., Goldstein, T., and Feizi, S. Curse of dimensionality on randomized smoothing for certifiable robustness. In *International Conference on Machine Learning*, pp. 5458–5467, 2020.

Li, B., Chen, C., Wang, W., and Carin, L. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems*, pp. 9464–9474, 2019.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2017.

Mendelson, S. and Vershynin, R. Entropy and the combinatorial dimension. *Inventiones mathematicae*, 152(1): 37–55, 2003.

Montasser, O., Hanneke, S., and Srebro, N. VC classes are adversarially robustly learnable, but only improperly. In *Annual Conference on Learning Theory*, pp. 2512–2530, 2019.

Northcutt, C. G., Athalye, A., and Mueller, J. Pervasive label errors in test sets destabilize machine learning benchmarks. In *NeurIPS 2021 Datasets and Benchmarks Track*, 2021.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, 2018.

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

Sun, J., Yang, Y., Xun, G., and Zhang, A. A stagewise hyperparameter scheduler to improve generalization. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1530–1540, 2021.

Sun, J., Huai, M., Jha, K., and Zhang, A. Demystify hyperparameters for stochastic optimization with transferable representations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1706–1716, 2022.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

von Luxburg, U. and Bousquet, O. Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, 2004.

Wu, X., Huang, F., Hu, Z., and Huang, H. Faster adaptive federated learning. *arXiv preprint arXiv:2212.00974*, 2022a.

Wu, X., Hu, Z., and Huang, H. Decentralized riemannian algorithm for nonconvex minimax problems. *arXiv preprint arXiv:2302.03825*, 2023.

Wu, Y., Bojchevski, A., Kuvshinov, A., and Günnemann, S. Completing the picture: Randomized smoothing suffers from the curse of dimensionality for a large family of distributions. In *International Conference on Artificial Intelligence and Statistics*, pp. 3763–3771. PMLR, 2021.

Wu, Y., Bojchevski, A., and Huang, H. Adversarial weight perturbation improves generalization in graph neural network. *arXiv preprint arXiv:2212.04983*, 2022b.

Wu, Y., Li, X., Kerschbaum, F., Huang, H., and Zhang, H. Towards robust dataset learning. *arXiv preprint arXiv:2211.10752*, 2022c.

Wu, Y., Zhang, H., and Huang, H. Retrievalguard: Provably robust 1-nearest neighbor image retrieval. In *International Conference on Machine Learning*, pp. 24266–24279. PMLR, 2022d.

Yang, G., Duan, T., Hu, J. E., Salman, H., Razenshteyn, I., and Li, J. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pp. 10693–10705, 2020a.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R., and Chaudhuri, K. A closer look at accuracy vs. robustness. In *Advances in Neural Information Processing Systems*, 2020b.

Yin, D., Kannan, R., and Bartlett, P. Rademacher complexity for adversarially robust generalization. In *International conference on machine learning*, pp. 7085–7094, 2019.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482, 2019.

A. Missing proofs

A.1. Proof of Lemma 3.2

Proof. Denote by X the random variable of μ on bounded space $\mathcal{X}$ We consider the $Z_1 = f(X)$ and $Z_0 = \mathbb{E}[f(X)]$, since

$$|Z_1 - Z_0| = |f(X) - \mathbb{E}[f(X)]| = |\mathbb{E}_{X'}[f(X) - f(X')]| \leq |L \sup_{x,x' \in \mathcal{X}} ||x - x'||| = L\text{diam}(\mathcal{X}),$$

where X' is of the same distribution with X. Because $\mathbb{E}[Z_1] = Z_0$, $\{Z_0, Z_1\}$ is a martingale with bounded difference. Thus, by Azuma's inequality Lemma 3.1, we have

$$\Pr(|f(x) - \mathbb{E}[f(x)]| \geq t) = \Pr(|Z_1 - Z_0| \geq t) \leq 2\exp(-\frac{t^2}{2\text{diam}(\mathcal{X})^2 L^2}).$$

□

A.2. Proof of Lemma 3.3

Proof. We use the similar proof technique as in Bubeck & Sellke (2023). Our proof depends on the following lemma.

Lemma A.1 (Lemma 2.1 of Bubeck & Sellke (2023))**.** *Let $\mathcal{F}$ be any class of functions from $\mathbb{R}^d \to [-1, 1]$. 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 $\sigma^2 := \mathbb{E}[\text{Var}[y|x]] > 0$.*

$$\Pr\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{8^3}\right) + \Pr\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n f(x_i)z_i \geq \frac{\epsilon}{4}\right).$$

We now try to bound the term $\Pr(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n f(x_i)z_i \geq \frac{\epsilon}{4})$. As x_i is randomly sampled from the input distribution and $\text{diam}(\mathcal{X}) = 2$, we have

$$\Pr(|f(x_i) - \mathbb{E}[f(x)]| \geq t) \leq 2\exp(-\frac{t^2}{8L^2}),$$

which indicates $f(x_i) - \mathbb{E}[f(x)]$ is $8L^2/n$-subgaussian distributed. Because $|z_i| = |y_i - g(x_i)| \leq 2$, we know $(f(x_i) - \mathbb{E}[f(x)])z_i$ is $32L^2$-subgaussian. By Property 1 in Case et al. (2019) we know $\frac{1}{n}\sum_{i=1}^n (f(x_i) - \mathbb{E}[f(x)])z_i$ is $32L^2/n$-subgaussian. Since $\mathbb{E}[(f(x_i) - \mathbb{E}[f(x)])z_i] = 0$, we have

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n (f(x_i) - \mathbb{E}[f(x)])z_i \geq \frac{\epsilon}{8}\right) \leq \exp(-\frac{n\epsilon^2}{2^{10}L^2}),$$

Since the range of the functions is in $[-1, 1]$ we have $\mathbb{E}[f(x)] \in [-1, 1]$ and hence:

$$\Pr\left(\exists f : \frac{1}{n}\sum_{i=1}^n \mathbb{E}[f(x)]z_i \geq \frac{\epsilon}{8}\right) \leq \Pr\left(|\frac{1}{n}\sum_{i=1}^n z_i| \geq \frac{\epsilon}{8}\right),$$

By Hoeffding's inequality, the above quantity is smaller than $2\exp(-n\epsilon^2/8^3)$ Thus we obtain with an union bound:

$$\begin{aligned}\Pr\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n f(x_i)z_i \geq \frac{\epsilon}{4}\right) &\leq |\mathcal{F}| \Pr\left(\frac{1}{n}\sum_{i=1}^n (f(x_i) - \mathbb{E}[f(x)])z_i \geq \frac{\epsilon}{8}\right) + \Pr\left(|\frac{1}{n}\sum_{i=1}^n z_i| \geq \frac{\epsilon}{8}\right) \\ &\leq |\mathcal{F}|\exp(-\frac{n\epsilon^2}{2^{10}L^2}) + 2\exp(-n\epsilon^2/8^3).\end{aligned}$$

Together with Lemma A.1 we have

$$\Pr\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon\right) \leq 4\exp\left(-\frac{n\epsilon^2}{8^3}\right) + |\mathcal{F}|\exp\left(-\frac{n\epsilon^2}{2^{10}L^2}\right),$$

which proves this lemma. □

A.3. Proof of Theorem 3.4

Proof. We use the similar proof technique as in Bubeck & Sellke (2023).

We argue that the η-covering of the function space $\mathcal{F}$ is upper bounded by the η/J-covering of the parameter space $\mathcal{W}$. To see this, we can select the centers $\mathcal{W}^c = \{w_i^c\}$ of the η/J-covering of $\mathcal{W}$, and covering $\mathcal{F}$ with η-balls centered at $f_{w_i^c}$, because $\forall f_w \in \mathcal{F}$, 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 $N(\mathcal{W}, \eta/J, ||\cdot||)$ balls. So we have

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

Taking $\eta = \frac{\epsilon}{6}$ and denote by $\mathcal{W}_\epsilon$ the $\epsilon/6J$-covering of the $\mathcal{W}$. Applying Lemma 3.3 to $\mathcal{F}_w = \{f_w : w \in \mathcal{W}_\epsilon\}$ we have

$$\Pr\left(\exists f \in \mathcal{F}_w : \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \leq \sigma^2 - \frac{\epsilon}{2} \text{ and } \mathrm{Lip}_{||\cdot||}(f) \leq L\right) \leq 4\exp\left(-\frac{n\epsilon^2}{8^3}\right) + \exp\left(p\ln(36JW\epsilon^{-1}) - \frac{n\epsilon^2}{2^{10}L^2}\right),$$

For all $f \in \mathcal{F}$, we can find an $f' \in \mathcal{F}_w$ such that $||f - f_w||_{\mathcal{F}} \leq \epsilon/6$. One can easily derive

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

Thus, if n is large enough such that $\exp(-n\epsilon^2/8^3) \leq \delta/8$ and $L \geq \frac{\epsilon}{32}\sqrt{\frac{n}{p\ln(36WJ\epsilon^{-1})+\ln(2/\delta)}}$, we have

$$\Pr\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \text{ and } \mathrm{Lip}_{||\cdot||}(f) \leq L\right) \leq \delta,$$

which yields with probability at least $1 - \delta$,

$$\frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \Rightarrow \mathrm{Lip}_{||\cdot||}(f) \geq \frac{\epsilon}{32}\sqrt{\frac{n}{p\ln(36WJ\epsilon^{-1}) + \ln(2/\delta)}}$$

□

A.4. Proof of Lemma 3.6

Proof. We consider the Lipschitz function class $B_L := \{f : \mathrm{Lip}_{||\cdot||}(f) \leq L\}$. 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 $\mathcal{U}_1, \mathcal{U}_2, ..., \mathcal{U}_N$ centered at $s_1, s_2, ..., s_N$. 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 $\widetilde{f}$ by taking its value on $\mathcal{U}_1$ as an $\epsilon/2$-approximation of $f(s_1)$. As $\mathrm{diam}(\mathcal{U}_1) \leq L \cdot \mathrm{diam}(\mathcal{X})$, there are at most $\lceil 2L \cdot \mathrm{diam}(\mathcal{X})/\epsilon \rceil$ 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 ϵ from $\widetilde{f}(s_1)$ by triangle inequality. It implies that to construct an ϵ-approximation of $f(s_2)$ on $\mathcal{U}_2$, we shall know either $\widetilde{f}(s_1) - \epsilon/2$ or $\widetilde{f}(s_1) + \epsilon/2$. Repeating the same argument by N times, we can bound the ϵ-covering of f on $\mathcal{X}$ by $\lceil 2L \cdot \mathrm{diam}(\mathcal{X})/\epsilon \rceil 2^N$. □

A.5. Proof of Lemma 3.7

Proof. The proof of this lemma is quite straight forward. Notice that when $u > 2L \cdot \mathrm{diam}(\mathcal{X})$, the number of u-covering for B_L is 1 and $\ln(N(B_L, u, ||\cdot||_{\mathcal{F}})) = 0$. Combining Equation 3 with Lemma 3.6 yields this lemma. □

A.6. Proof of Lemma 3.8

Proof. As $\frac{u}{2L} \leq \mathrm{diam}(\mathcal{X})$, we have $N(\mathcal{X}, \frac{u}{2L}, ||\cdot||) \leq (\frac{12L}{u})^d$ and

$$\begin{aligned}
\mathbb{E}_{S\in\mathcal{D}^n}[R(B_L \circ S)] &\leq 2\epsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\epsilon/4}^{2L\cdot\mathrm{diam}(\mathcal{X})} \sqrt{N\left(\mathcal{X}, \frac{u}{2L}, ||\cdot||\right) \ln 2 + \ln\left(\left\lceil \frac{2L\cdot\mathrm{diam}(\mathcal{X})}{u} \right\rceil\right)}\, du \\
&\leq 2\epsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\epsilon/4}^{4L} \sqrt{\left(\frac{12L}{u}\right)^d \ln 2 + \ln\left(\left\lceil \frac{2L}{u} \right\rceil\right)}\, du \\
&\leq 2\epsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\epsilon/4}^{4L} \left[\sqrt{\left(\frac{12L}{u}\right)^d \ln 2} + \sqrt{\ln\left(\left\lceil \frac{2L}{u} \right\rceil\right)}\right] du \\
&\leq 2\epsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\epsilon/4}^{4L} \sqrt{\left(\frac{12L}{u}\right)^d \ln 2}\, du + \frac{16\sqrt{2}L}{\sqrt{n}} \sqrt{\ln(16L/\epsilon + 1)}.
\end{aligned}$$

Switching the integral variable from u to $v = u/12L$ we have

$$\begin{aligned}
\int_{\epsilon/4}^{4L} \sqrt{\left(\frac{12L}{u}\right)^d \ln 2}\, du &= 12L \int_{\epsilon/(48L)}^{1/3} \sqrt{v^{-d} \ln 2}\, dv \\
&= 12L \left[\sqrt{\ln 2} \frac{1}{-d/2+1} v^{-d/2+1} |_{\epsilon/(48L)}^{1/3})\right] \\
&< 12L \frac{2\sqrt{\ln 2}}{d-2} \left(\frac{48L}{\epsilon}\right)^{d/2-1}.
\end{aligned}$$

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

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

□

A.7. Proof of Theorem 3.9

Proof. According to Equation 1,

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

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,

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

Under the label noise settings, we have

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}}(f) &= \mathbb{E}_{\mathcal{D}}[(f(x) - y)^2] \\
&= \mathbb{E}_{x,y}[(f(x) - \mathbb{E}_y[y|x])^2 + (y - \mathbb{E}_y[y|x])^2] \\
&\geq \mathbb{E}_x[\mathrm{Var}(y|x)] = \sigma^2.
\end{aligned}$$

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

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

Consider $B_L = \{f \in \mathcal{F} : \mathrm{Lip}_{||\cdot||}(f) \leq L\}$. According to Lemma 3.8, 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 $K = 96 + \frac{96\sqrt{2\ln 2}}{d-2} + \frac{16\sqrt{2}}{n^{1/2-1/d}}\sqrt{\ln(\frac{1}{3}n^{1/d}+1)} \sim \Theta(1)$. Thus we have

$$L \geq \frac{n^{1/d}}{K}\left(\frac{1}{8}\epsilon - \frac{1}{2}\sqrt{\frac{2\ln(2/\delta)}{n}}\right).$$

If $\exists f_0 \in \mathcal{F}$, such that

$$\mathcal{L}_S(f_0) \leq \sigma^2 - \epsilon \Rightarrow \mathrm{Lip}_{\|\cdot\|}(f_0) < \frac{n^{1/d}}{K}\left(\frac{1}{8}\epsilon - \frac{1}{2}\sqrt{\frac{2\ln(2/\delta)}{n}}\right),$$

we have

$$\begin{aligned}&\frac{\epsilon}{8} - \frac{1}{2}\sqrt{\frac{2\ln(2/\delta)}{n}} > K\frac{\mathrm{Lip}_{\|\cdot\|}(f_0)}{n^{1/d}} \geq \\ &\mathbb{E}_{S\in\mathcal{D}^n}[R(B_{\mathrm{Lip}_{\|\cdot\|}(f_0)}\circ S)] \geq \frac{\epsilon}{8} - \frac{1}{2}\sqrt{\frac{2\ln(2/\delta)}{n}},\end{aligned}$$

which yields contradiction. Therefore, $\forall f \in \mathcal{F}$,

$$\mathcal{L}_S(f) \leq \sigma^2 - \epsilon \Rightarrow \mathrm{Lip}_{\|\cdot\|}(f) \geq \frac{n^{1/d}}{K}\left(\frac{1}{8}\epsilon - \frac{1}{2}\sqrt{\frac{2\ln(2/\delta)}{n}}\right).$$

Taking $\mathcal{X} = \{x \in \mathbb{R}^d : ||x|| \leq 1\}$, we have $\mathrm{diam}(\mathcal{X}) = 2$, which yields Theorem 3.9. □

A.8. Proof of Theorem 3.11

Proof. First, we show that we can find n training samples $\{x_1, ..., x_n\}$ such that $\forall i, j, i \neq j, ||x_i - x_j|| \geq \frac{1}{n^{1/d}}$. Consider the $\frac{1}{n^{1/d}}$-packing of the space $\{x : ||x|| \leq 1\}$, the packing number is greater than the $\frac{1}{n^{1/d}}$-covering number of the same space, which at least $(1/\frac{1}{n^{1/d}})^d = n$, we then choose $\{x_1, ..., x_n\}$ from the $\frac{1}{n^{1/d}}$-packing, the minimum pairwise distance is at least $\frac{1}{n^{1/d}}$. 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 $\frac{|y_i - y_j|}{||x_i - x_j||} \leq 2n^{1/d}$. □

A.9. Proof of Lemma 4.1

Proof. Our proof is partly based on Theorem 5.1 of Shalev-Shwartz & Ben-David (2014). Let $\mathcal{C}$ be a subset of $\mathcal{X}$ of size $2m$. There exist $T = 2^{2m}$ possible labeling functions from $\mathcal{C}$ to $\{-a, a\}$. Denote these functions by $f_1, ..., f_T$. We then define a distribution $\mathcal{D}_i$ w.r.t. f_i by

$$\mathcal{D}_i(\{(x,y)\}) = \begin{cases} p/|\mathcal{C}|, & \text{if } y = f_i(x); \\ (1-p)/|\mathcal{C}|, & \text{if } y \neq f_i(x), \end{cases}$$

where $p > 1/2$ satisfies $\mathrm{Var}(y|x) = \sigma^2 = 4a^2p(1-p)$ (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 $\mathcal{C} \times \{-a, a\}$ and returns a function $\mathcal{A}(S) : \mathcal{C} \to \mathbb{R}$, it holds that

$$\max_{i \in [T]} \mathbb{E}_{S \sim \mathcal{D}_i^m}[\mathcal{L}_{\mathcal{D}_i}(\mathcal{A}(S))] \geq \frac{a^2 + \sigma^2}{2}.$$

There are $k = (2m)^m$ possible sequences of m instances from $\mathcal{C}$. Denote these sequences by $S_1, ..., S_k$. Also, if $S_j = (x_1, ..., x_m)$, we denote by S_j^i the sequence containing the instances in S_j labeled by the function f_i, namely, $S_j^i = ((x_1, a_1 f_i(x_1)), ..., (x_m, a_m f_i(x_m)))$, where $\Pr(a_l = 1) = p$, $\Pr(a_l = -1) = 1 - p$, and $a_1, ..., a_m$ 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 $S_1^i, ..., S_k^i$, and all these training sets have the same probability of being sampled. Therefore,

$$\mathbb{E}_{S \sim \mathcal{D}_i^m}[\mathcal{L}_{\mathcal{D}_i}(\mathcal{A}(S))] = \frac{1}{k} \sum_{j=1}^{k} \mathcal{L}_{\mathcal{D}_i}(\mathcal{A}(S_j^i)).$$

Using the facts that "maximum" is larger than "average" and that "average" is larger than "minimum", we have

$$\begin{aligned}
\max_{i \in [T]} \frac{1}{k} \sum_{j=1}^{k} \mathcal{L}_{\mathcal{D}_i}(\mathcal{A}(S_j^i)) &\geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{k} \sum_{j=1}^{k} \mathcal{L}_{\mathcal{D}_i}(\mathcal{A}(S_j^i)) \\
&= \frac{1}{k} \sum_{j=1}^{k} \frac{1}{T} \sum_{i=1}^{T} \mathcal{L}_{\mathcal{D}_i}(\mathcal{A}(S_j^i)) \\
&\geq \min_{j \in [k]} \frac{1}{T} \sum_{i=1}^{T} \mathcal{L}_{\mathcal{D}_i}(\mathcal{A}(S_j^i)).
\end{aligned}$$

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

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

Note that

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

Hence,

$$\begin{aligned}
\frac{1}{T} \sum_{i=1}^{T} \mathcal{L}_{\mathcal{D}_i}(\mathcal{A}(S_j^i)) &\geq \frac{1}{T} \sum_{i=1}^{T} \mathbb{E}_{a \in \{-1,1\}^m} \left[\sigma^2 + \frac{1}{2q} \sum_{r=1}^{q} (\mathcal{A}(S_j^i(a))(v_r) - (2p-1) f_i(v_r))^2 \right] \\
&= \sigma^2 + \frac{1}{2q} \sum_{r=1}^{q} \frac{1}{T} \sum_{i=1}^{T} \mathbb{E}_{a \in \{-1,1\}^m} [(\mathcal{A}(S_j^i(a))(v_r) - (2p-1) f_i(v_r))^2] \\
&\geq \sigma^2 + \frac{1}{2} \min_{r \in [p]} \frac{1}{T} \sum_{i=1}^{T} \mathbb{E}_{a \in \{-1,1\}^m} [(\mathcal{A}(S_j^i)(a)(v_r) - (2p-1) f_i(v_r))^2].
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}_{a\in\{-1,1\}^m}[(\mathcal{A}(S_j^i)(v_r) - (2p-1)f_i(v_r))^2] + \mathbb{E}_{a\in\{-1,1\}^m}[(\mathcal{A}(S_j^{i'})(v_r) - (2p-1)f_{i'}(v_r))^2] \\
\geq &\mathbb{E}_{a\in\{-1,1\}^m}[(\mathcal{A}(S_j^i)(v_r) - (2p-1)f_i(v_r))^2 + (\mathcal{A}(S_j^{i'})(v_r) - (2p-1)f_{i'}(v_r))^2] \\
\geq &\mathbb{E}_{a\in\{-1,1\}^m}\left[\frac{1}{2}(2p-1)^2(f_{i'}(v_r) - f_i(v_r))^2\right] \\
= &2(2p-1)^2a^2,
\end{aligned}$$

which yields

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

Combining the discussion above, we have

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

□

A.10. Proof of Lemma 4.2

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 $(x_1, x_2) = \operatorname{argmin}_{(a,b)\in\mathcal{C}\times\mathcal{C}, a\neq b}||a-b||$. We can select $\mathcal{D}$ such that $\mathcal{D}(\{(x_1,1)\}) = \frac{p}{2}, \mathcal{D}(\{(x_1,-1)\}) = \frac{(1-p)}{2}$ and $\mathcal{D}(\{(x_2,-1)\}) = \frac{p}{2}, \mathcal{D}(\{(x_2,-1)\}) = \frac{1-p}{2}$, where $4p(1-p) = \sigma^2, p > 1/2$. Consider an L-Lipschitz learning algorithm $\mathcal{A}(S) : \mathcal{C} \to \mathbb{R}$:

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

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 **??**, for every $\mathcal{A}(S) : \mathcal{X} \to \mathbb{R}$ and every $\mathcal{C} \subseteq \mathcal{X}$ there exists $\mathcal{D}$ such that

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

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

$$n < \max_{\mathcal{C}\subseteq\mathcal{X}, d(\mathcal{C})\geq\epsilon}\frac{|\mathcal{C}|}{2} \;\Rightarrow\; \mathbb{E}_{S\sim\mathcal{D}^n}[\mathcal{L}_{\mathcal{D}}(\mathcal{A}(S))] \geq \frac{1+\sigma^2}{2}.$$

Denote the ϵ-packing number of space $(\mathcal{X}, ||\cdot||)$ by $M(\mathcal{X}, \epsilon, ||\cdot||)$. We have

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

That is,

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

Combining a) and b) yields our conclusion. □

A.11. Proof of Theorem 4.3

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 $\sigma^2 \leq 0.5$

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

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

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

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

B. Some basic concepts of Rademacher complexity

Definition B.1 (Representativeness of S)**.**

$$Rep_{\mathcal{D}}(l, \mathcal{F}, S) := \sup_{f\in\mathcal{F}}(\mathcal{L}_D(f) - \mathcal{L}_S(f)).$$

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

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

Lemma B.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 $f \in \mathcal{F}$,*

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

Lemma B.4 (Lemma 26.2 in Shalev-Shwartz & Ben-David (2014))**.**

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

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

Lemma B.5 (Theorem 26.5 in Shalev-Shwartz & Ben-David (2014))**.** *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 $f \in \mathcal{F}$,*

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

Lemma B.6 (Lemma 26.9 in Shalev-Shwartz & Ben-David (2014))**.** *If $l(f(x), y)$ is $C_{||\cdot||}$-Lipschitz w.r.t. $f(x)$ for arbitrary $y \in [-1, 1]$,*

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