Statistica Sinica Preprint No: SS-2024-0298	
Title	Minimax Rates of Convergence for Nonparametric Regression Under Adversarial Attacks
Manuscript ID	SS-2024-0298
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0298
Complete List of Authors	Jingfu Peng and Yuhong Yang
Corresponding Authors	Yuhong Yang
E-mails	yangx374@umn.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

Minimax Rates of Convergence for Nonparametric Regression Under Adversarial Attacks

Jingfu Peng[1] and Yuhong Yang[1,2]

[1] *Yau Mathematical Sciences Center, Tsinghua University*

[2] *Beijing Institute of Mathematical Sciences and Applications*

Abstract: Recent research shows the susceptibility of machine learning models to adversarial attacks, wherein minor but maliciously chosen perturbations of the input can significantly degrade model performance. In this paper, we theoretically analyse the limits of robustness against such adversarial attacks in a nonparametric regression setting, by examining the minimax rates of convergence in an adversarial sup-norm. Our work reveals that the minimax rate under adversarial attacks in the input is the same as sum of two terms: one represents the minimax rate in the standard setting without adversarial attacks, and the other reflects the maximum deviation of the true regression function value within the target function class when subjected to the input perturbations. The optimal rates under the adversarial setup can be achieved by an adversarial plug-in procedure constructed from a minimax optimal estimator in the corresponding standard setting. Two specific examples are given to illustrate the established minimax results.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

Key words and phrases: Nonparametric regression, adversarial robustness, minimax risk, sup-norm.

1. Introduction

Over the past decade, machine/deep learning models have found unprecedented applications in a variety of domains including image recognition (Krizhevsky et al., 2012), natural language and speech processing (Collobert et al., 2011), game playing (Silver et al., 2016), autonomous driving (Grigorescu et al., 2020), many of which are safety-critical. However, it is found that these learning models are vulnerable to adversarial attacks. Here, an adversary is able to change the inputs to an already trained model, but cannot modify the training process. For example, input perturbations due to changes of weather conditions can significantly degrade the accuracy of trained neural networks for traffic sign recognition, demonstrating that such natural input variations present a significant challenge for deep learning (Robey et al., 2020). Besides the nature as an adversary, a malicious opponent may choose perturbations to maximize prediction errors of a well trained neural network model (Szegedy et al., 2013). Similar vulnerabilities have been observed in various models across different application areas (see, e.g., Biggio et al., 2013; Goodfellow et al., 2014; Papernot et al., 2016).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

The concerns about the safety and reliability of machine learning models have motivated a growing body of research focused on crafting the adversarial examples (Goodfellow et al., 2014; Papernot et al., 2016; Moosavi-Dezfooli et al., 2016; Carlini and Wagner, 2017; Awasthi et al., 2020) and devising defenses to enhance model robustness against such perturbations (Goodfellow et al., 2014; Madry et al., 2017; Finlay and Oberman, 2021; Raghunathan et al., 2018; Cohen et al., 2019). Adversarial training, which minimizes the empirical risk under worst-case perturbations on the training data, has been empirically demonstrated to be effective against various attacks (see, e.g., Madry et al., 2017). While considerable efforts have been made on constructing attack and defence, the problem of understanding the intrinsic hardness of estimation and assessing the optimality of learning methods under adversarial attacks are far less understood.

One of the most important approach to measuring the difficulty of a nonparametric statistical problem is to evaluate its minimax risk (see, e.g., Ibragimov and Khas' minskii, 1982; Birgé, 1986; Yang and Barron, 1999). In the adversarial setting, the maximal risk of an estimator is defined as its worst statistical performance over a class of distributions when the input perturbation is generated from a given perturbation set to deprave the model's performance. If its maximal risk is minimal (rate) among all esti-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

mators, then this estimator is called minimax (rate) optimal. To the best of our knowledge, investigating the adversarial robustness from the minimax viewpoint has not been paid much attention. Dan et al. (2020) considered a binary classification problem with data generated from a Gaussian mixture model. They established the minimax rate of excess risk when the perturbations lie in an origin-symmetric convex set. Xing et al. (2021) determined the minimax rate of a nonparametric classification problem when the testing input is randomly perturbed on a sphere, and established the minimax optimality of a nearest neighbor rule. In a setup of linear regression with Gaussian regressors, Xing et al. (2021) provided the minimax rate for estimating regression coefficients under bounded ℓ_2-norm perturbations. In a context of data contamination where a subset of training sample can be arbitrarily modified by an attacker, Zhao and Wan (2024) established the minimax rates for the estimation of a nonparametric Lipschitz regression function under both ℓ_2 and ℓ_∞ losses. Although the above theoretical advancements provide valuable insights, they are confined to some restricted setups based on simple models and architectures, and thus do not seem to be applicable to the broader nonparametric setting with the adversarial attacks as we consider.

Under a nonparametric regression setting with minimal assumptions re-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

garding the adversarial perturbations, an important question arises: What is the minimax rate of convergence for a general class of regression functions?

This paper determines the sup-norm rate of convergence in a nonparametric regression setup with additive perturbations, in which the attacker can add arbitrary perturbations in a set to the input, thereby degrading the performance of the trained estimator. We establish that under general class of regression functions and adversarial perturbation sets, the minimax risk converges at the order of the rate in the standard setup without adversaries, plus the maximum deviation of true function values within the target regression function class. The optimal rate can be achieved by an adversarial plug-in procedure constructed from a minimax optimal estimator in the standard setting. We provide minimax results for two specific examples of function classes, including isotropic Hölder class and anisotropic Hölder class, and investigate the effects of ℓ_p-attacks ($0 < p \leq \infty$) and sparse attacks under these two function classes, respectively.

1.1 Related work

Sup-norm convergence. Determining the rate of convergence in the sup-norm is a crucial topic in statistics and machine learning. Classical contri-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

butions include works by Devroye (1978); Stone (1982); Donoho (1994); Korostelev and Nussbaum (1999); Lepski and Tsybakov (2000); Bertin (2004a); Gaïffas (2007); Giné and Nickl (2009); Chen and Christensen (2015). More recently, the implications of sup-norm convergence in transfer learning have been explored by Schmidt-Hieber and Zamolodtchikov (2024), and its relation to adversarial training has been investigated by Imaizumi (2023). However, these studies focus on standard setups without adversarial perturbations to the input data.

Robustness of nonparametric classifiers. Several previous works analyzed the robustness of specific families of classifiers. Wang et al. (2018) studied the robustness of nearest neighbor classifier. Yang et al. (2020) proposed the attack strategies that apply to a wide range of non-parametric classifiers and analyzed a general defense method based on data pruning. Bhattacharjee and Chaudhuri (2020) proved the consistency of the nearest neighbor and kernel estimators. Note that the aforementioned works do not establish the optimal rate of convergence of nonparametric estimation under the adversarial attacks.

Distributional robustness optimization. Lee and Raginsky (2018) and Tu et al. (2019) established the connections between the adversarial training and distributional robustness optimization (DRO) (Ben-Tal et al., 2009;

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

Shapiro et al., 2021). These connections can be used to upper bound the generalization error of the adversarial training. In the context of DRO, when the loss function is defined as a product of the response variable and the parameter, Duchi et al. (2023) obtained minimax lower bounds for a distributionally robust loss. However, the linear form of the loss function in their work cannot be applied to the typical regression setting.

Other related work. Rather than studying the minimax risk, another line of work obtained tight statistical characterizations of the Bayes adversarial risk and developed classifiers to realized it (Schmidt et al., 2018; Bhagoji et al., 2019; Pydi and Jog, 2020). The trade-offs between standard and robust accuracy have been studied by Madry et al. (2017); Schmidt et al. (2018); Tsipras et al. (2019); Raghunathan et al. (2019); Zhang et al. (2019); Javanmard et al. (2020); Min et al. (2021); Mehrabi et al. (2021); Dobriban et al. (2020); Javanmard and Soltanolkotabi (2022). Algorithm-free generalization bounds such as VC-dimension have been studied by Attias et al. (2019); Montasser et al. (2019) in the adversarial setting. Rademacher complexity of the adversarial training has been investigated by Yin et al. (2019); Khim and Loh (2018); Awasthi et al. (2020). Recently, Liu et al. (2023) derived non-asymptotic bounds for adversarial excess risk under misspecified models. Note that the above analyses primarily center on upper

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

bounding the adversarial risk, thus lacking corresponding lower bounds necessary for determining the minimax rates.

1.2 Outline

The rest of this paper is organized as follows. Section 2 gives a setup for the nonparametric regression problem and the definition of adversarial loss/risk. In Section 3, we state upper and lower bounds on the minimax risks under the adversarial attack. Two specific examples are discussed in Section 4. Section 5 presents numerical simulation results. The proofs of the main theorems and examples are provided in the supplementary material.

2. Problem setup

This paper considers the problem of nonparametric regression estimation. Suppose the observations $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ are generated from the regression model

$$Y_i = f(X_i) + \xi_i, \tag{2.1}$$

where $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{Y} \subseteq \mathbb{R}$, $f : \mathcal{X} \to \mathcal{Y}$ is an unknown regression function, ξ_i is a random error term with $\mathbb{E}(\xi_i|X_i) = 0$ a.s., and X_i follows an unknown marginal distribution $\mathbb{P}_X$ on $\mathcal{X}$. The goal is to develop an estimator $\hat{f}$ of f based on the observed data. The estimation accuracy of $\hat{f}$ is measured by

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

the sup-norm loss. In the standard setting of regression with unperturbed future X values, this loss is defined as $\sup_{x\in\mathcal{X}} |f(x)-\hat{f}(x)|$, which quantifies the uniform convergence of $\hat{f}$ to f over $\mathcal{X}$.

In this paper, we consider the estimation of the regression function in the presence of an adversary. Specifically, when assessing the performance of the estimator $\hat{f}$, the adversary can add any perturbation $\delta \in \Delta_n$ to the input x, where $\Delta_n \in \mathbb{R}^d$ is a closed set containing $\delta = 0$, and Δ_n may depend on the sample size n. A representative example of Δ_n is the ℓ_p-ball $B_p^{q_n} = \{z : \|z\|_p \leq q_n\}$ centering at origin with radius $q_n > 0$ and $p > 0$. In the adversarial setting, the sup-norm loss of estimation is defined as

$$L_{\Delta_n}(f, \hat{f}) = \sup_{x\in\mathcal{X}} \sup_{\substack{\delta\in\Delta_n \\ x+\delta\in\mathcal{X}}} \left| f(x) - \hat{f}(x+\delta) \right|, \tag{2.2}$$

and the corresponding adversarial risk is given by

$$R_{\Delta_n}(f, \hat{f}) = \mathbb{E} L_{\Delta_n}(f, \hat{f}), \tag{2.3}$$

where the expectation $\mathbb{E}$ is taken with respect to the observed data generated from the regression model (2.1), and the subscript Δ_n here is employed to emphasize the dependence of the adversarial risk/loss on the perturbation set Δ_n. In the standard regression setting with $\Delta_n = \{0\}$, expressions (2.2) and (2.3) reduce to the standard sup-norm loss

$$L(f, \hat{f}) = \sup_{x\in\mathcal{X}} \left| f(x) - \hat{f}(x) \right|$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

and the standard sup-norm risk

$$R(f, \hat{f}) = \mathbb{E}L(f, \hat{f}),$$

respectively. In the adversarial setting, an estimator $\hat{f}$ is sought to be robust to the adversarial perturbation of x.

The regression function f is assumed to belong to a function class $\mathcal{F}$. The minimax risk of estimating $f \in \mathcal{F}$ under the adversarial sup-norm loss is expressed as:

$$V_{\Delta_n} = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} R_{\Delta_n}(f, \hat{f}). \tag{2.4}$$

Then two important questions arise:

Q1. What factors determine the rate of convergence of V_{Δ_n}?

Q2. How can minimax optimal procedures be developed to achieve the optimal rate of V_{Δ_n}?

Answers to questions Q1 and Q2 have the potential to offer previously unavailable insights into the theoretical foundations and practical applications of adversarial learning.

Throughout this paper, let $\mathbb{N}_0$ denote the set of non-negative integers. For any $a \in \mathbb{R}^d$ and $B \subseteq \mathbb{R}^d$, we use the Minkowski sum notations $a + B \triangleq \{a + b : b \in B\}$ and $a - B \triangleq \{a - b : b \in B\}$. For any positive sequences

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

a_n and b_n, we denote $a_n = O(b_n)$ and $a_n \lesssim b_n$ if there exist $C > 0$ and $N > 0$ such that $n \geq N$ implies $a_n \leq C b_n$. If $a_n = O(b_n)$ and $b_n = O(a_n)$, then we write $a_n \asymp b_n$. For $1 \leq p < \infty$, we use $\|\delta\|_p$ to denote the ℓ_p-norm $(\sum_{j=1}^d |\delta|_j^p)^{1/p}$ of a vector $\delta \in \mathbb{R}^d$. We use $\|\delta\|_\infty$ to denote the sup-norm $\sup_{1\leq j\leq d} |\delta_j|$. For brevity, we write $\|\delta\|$ to represent the ℓ_2-norm.

3. Main results

In this section, we begin by deriving a closed form expression for the ideal adversarial loss $\inf_{f'} L_{\Delta_n}(f, f')$. Then we establish the minimax rates of convergence for the general function classes $\mathcal{F}$ and perturbation sets Δ_n.

3.1 Ideal adversarial loss

We first introduce an equivalent form for the adversarial sup-norm loss (2.2), which offers conveniences in characterizing both the ideal adversarial loss and the minimax risk V_{Δ_n}.

Lemma 1. *For any estimator $\hat{f}$, we have*

$$L_{\Delta_n}(f, \hat{f}) = \sup_{x\in\mathcal{X}} \sup_{x'\in(x+\Delta_n)\cap\mathcal{X}} \left|f(x) - \hat{f}(x')\right| = \sup_{x'\in\mathcal{X}} \sup_{x\in(x'-\Delta_n)\cap\mathcal{X}} \left|f(x) - \hat{f}(x')\right|. \tag{3.1}$$

Lemma 1 provides an alternative expression for the adversarial loss by exchanging the order of two supremum operations. The inner supremum

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

in the last argument of (3.1), which depends on the perturbation set, is taken respect to the regression function f rather than the estimator $\hat{f}$. This property facilitates the derivation of the ideal adversarial loss and the ideal adversarial estimator (i.e., the best performing "estimate" when the underlying regression function f is known). The next theorem addresses this aspect.

Theorem 1. *Given the regression function f, the ideal adversarial loss is given by*

$$L^*_{\Delta_n}(f) \triangleq \inf_{f'} L_{\Delta_n}(f, f') = \frac{1}{2} \sup_{x' \in \mathcal{X}} \left\{ \sup_{x \in (x' - \Delta_n) \cap \mathcal{X}} f(x) - \inf_{x \in (x' - \Delta_n) \cap \mathcal{X}} f(x) \right\}, \tag{3.2}$$

where the minimum is achieved by the adversarial regression function:

$$f^*(x) = \frac{1}{2} \left\{ \sup_{x' \in (x - \Delta_n) \cap \mathcal{X}} f(x') + \inf_{x' \in (x - \Delta_n) \cap \mathcal{X}} f(x') \right\}, \quad x \in \mathcal{X}. \tag{3.3}$$

Theorem 1 provides a closed form expression for the ideal adversarial loss, which shows that the ideal adversarial loss is proportional to the maximum variation of the true regression function value within the perturbation set Δ_n over the domain $\mathcal{X}$. Moreover, the ideal adversarial regression function is exactly the average of the maximum and minimum values of the function f in the adversarial neighborhood $(x - \Delta_n) \cap \mathcal{X}$.

The result from Theorem 1 substantiates that the optimal adversarial

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

robustness is jointly determined by the size of the perturbation set and the smoothness of the true regression function. For example, when f satisfies the Lipschitz smoothness condition $|f(x) - f(z)| \leq L \cdot \|x - z\|$ and Δ_n has the diameter $\mathrm{diag}(\Delta_n) \triangleq \max_{x,z} \|x - z\|$, then the ideal adversarial loss

$$L^*_{\Delta_n}(f) \leq \frac{L \cdot \mathrm{diag}(\Delta_n)}{2},$$

a quantity controllable when the diameter of Δ_n is not excessively large. In contrast, if the true regression function is discontinuous, then $L^*_{\Delta_n}(f)$ cannot degenerate to 0 unless $\Delta_n = \{0\}$. Also, if Δ_n does not shrink with n, $L^*_{\Delta_n}(f)$ may not converge to 0.

Remark 1. In the literature, several papers have obtained precise characterizations or tight bounds on the ideal adversarial loss (see, e.g., Bhagoji et al., 2019; Pydi and Jog, 2020; Dan et al., 2020; Xing et al., 2021). However, it is important to note that all of these works focus on parametric models, which cannot imply the adversarial robustness for nonparametric regression as considered in this paper.

3.2 Minimax rates of convergence

In this subsection, our aim is to establish the minimax rates of convergence for the sup-norm risk under the adversarial attacks. We propose an adversarial plug-in procedure to achieve the minimax optimal rates, which is

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

derived from a minimax optimal estimator in the corresponding standard setting.

In Theorem 1, we obtain the explicit expression for the ideal adversarial regression function (3.3). However, (3.3) is infeasible in practice as it relies on the true regression function f. Motivated by (3.3), we devise a feasible adversarial estimator through the following two steps:

Step 1. Utilizing the observed data $(X_1, Y_1), \ldots, (X_n, Y_n)$, we construct an estimator $\tilde{f}$ for the regression function f.

Step 2. Subsequently, we formulate an adversarial plug-in estimator:

$$\hat{f}_{\mathrm{PI}}(x) = \frac{1}{2}\left\{ \sup_{x' \in (x-\Delta_n)\cap\mathcal{X}} \tilde{f}(x') + \inf_{x' \in (x-\Delta_n)\cap\mathcal{X}} \tilde{f}(x') \right\}, \quad x \in \mathcal{X}. \tag{3.4}$$

The performance of the adversarial plug-in estimator $\hat{f}_{\mathrm{PI}}(x)$ clearly depends on the construction of $\tilde{f}$. The following theorem first provides an upper bound for the adversarial risk of $\hat{f}_{\mathrm{PI}}(x)$ considering a general $\tilde{f}$. Additionally, Theorem 2 establishes minimax upper bounds when specific choices of $\tilde{f}$ are adopted.

Theorem 2 (Upper bound)**.** *For any regression function f and any estimator $\tilde{f}$, the adversarial risk of the plug-in estimator (3.4) is upper bounded by*

$$R_{\Delta_n}(f, \hat{f}_{\mathrm{PI}}) \leq R(f, \tilde{f}) + L^*_{\Delta_n}(f), \tag{3.5}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

where $L^*_{\Delta_n}(f)$ *is the ideal adversarial loss defined in (3.2).*

Moreover, given a function class $\mathcal{F}$*, if* $\tilde{f}$ *satisfies*

$$\sup_{f\in\mathcal{F}} R(f,\tilde{f}) \asymp \inf_{\hat{f}} \sup_{f\in\mathcal{F}} R(f,\hat{f}), \tag{3.6}$$

then the adversarial maximal risk of $\hat{f}_{\mathrm{PI}}$ *is upper bounded by*

$$\sup_{f\in\mathcal{F}} R_{\Delta_n}(f,\hat{f}_{\mathrm{PI}}) \lesssim \inf_{\hat{f}} \sup_{f\in\mathcal{F}} R(f,\hat{f}) + \sup_{f\in\mathcal{F}} L^*_{\Delta_n}(f). \tag{3.7}$$

The relationship (3.5) illustrates that the adversarial risk of the plug-in estimator $\hat{f}_{\mathrm{PI}}$ can be upper bounded by the standard risk of the original estimator $\tilde{f}$ plus a multiple of the ideal adversarial loss $L^*_{\Delta_n}(f)$. Importantly, this relation holds without any additional constraints on the true regression function and the perturbation set, and without imposing assumptions on the estimator $\tilde{f}$. The second part of Theorem 2 indicates that if the original estimator $\tilde{f}$ is minimax optimal in the standard setting, then the corresponding adversarial maximal risk $\sup_{f\in\mathcal{F}} R_{\Delta_n}(f,\hat{f}_{\mathrm{PI}})$ is upper bounded by the standard minimax rate plus $\sup_{f\in\mathcal{F}} L^*_{\Delta_n}(f)$.

The following lower bound results show that the adversarial plug-in estimator based on $\tilde{f}$ with (3.6) is in fact minimax rate optimal.

Theorem 3 (Lower bound)**.** *For any regression function* f *and any estimator* $\hat{f}$*, the adversarial risk is lower bounded by*

$$R_{\Delta_n}(f,\hat{f}) \geq R(f,\hat{f}) \vee L^*_{\Delta_n}(f). \tag{3.8}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

Furthermore, for any function class $\mathcal{F}$, we have

$$\inf_{\hat{f}} \sup_{f\in\mathcal{F}} R_{\Delta_n}(f,\hat{f}) \gtrsim \inf_{\hat{f}} \sup_{f\in\mathcal{F}} R(f,\hat{f}) + \sup_{f\in\mathcal{F}} L^*_{\Delta_n}(f). \tag{3.9}$$

In summary, Theorems 2–3 together establish the minimax rates of convergence for nonparametric regression under the adversarial attacks,

$$\inf_{\hat{f}} \sup_{f\in\mathcal{F}} R_{\Delta_n}(f,\hat{f}) \asymp \inf_{\hat{f}} \sup_{f\in\mathcal{F}} R(f,\hat{f}) + \sup_{f\in\mathcal{F}} L^*_{\Delta_n}(f). \tag{3.10}$$

Therefore, (3.10) addresses Question Q.1 raised in Section 2, showing that the adversarial minimax rate is jointly determined by the standard minimax rate and the largest ideal loss in $\mathcal{F}$. Regarding Question Q.2, we establish that if $\tilde{f}$ is minimax optimal in the sense that $\sup_{f\in\mathcal{F}} R(f,\tilde{f}) \asymp \inf_{\hat{f}} \sup_{f\in\mathcal{F}} R(f,\hat{f})$ under the standard setting, then the adversarial plug-in estimator $\hat{f}_{\mathrm{PI}}$ based on $\tilde{f}$ is minimax optimal in terms of the adversarial risk. To the best our knowledge, (3.10) is the first minimax result in adversarial learning for the general regression setting. Our bounds are modular and can be applied to many models by computing the sup-norm convergence and the ideal adversarial loss in the target function class.

4. Applications

In this section, we demonstrate the applications of the theorems in the previous section through specific examples of function classes and perturbation

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

sets. We consider the case $\mathcal{X} = [0,1]^d$, and $(X_1, Y_1), \ldots, (X_n, Y_n)$ are drawn i.i.d. according to the regression model (2.1). The following assumption on the distribution of X is required.

Assumption 1. The marginal distribution $\mathbb{P}_X$ admits a density function that is lower bounded away from 0 and upper bounded by a positive constant on $\mathcal{X}$.

Assumption 1 ensures that the covariates X are more or less evenly distributed over the compact support $[0,1]^d$. As a result, there are sufficiently many observations around any point in the support, allowing for the construction of well-behaved estimators for the regression function in the sup-norm loss. This assumption is standard in nonparametric regression with random design; see, for example, Condition 3' in Stone (1982) and Definition 2.2 in Audibert and Tsybakov (2007). In addition, we further assume that the random error term is distributed according to a centered Gaussian distribution, which is the scenario where the known minimax theory in sup-norm can apply (see, e.g., Stone, 1982; Bertin, 2004b; Gaïffas, 2007).

Assumption 2. The random error term ξ follows a zero-mean Gaussian distribution and is independent of X.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

4.1 Isotropic Hölder class

Let $\beta = k + \alpha$ for some $k \in \mathbb{N}_0$ and $0 < \alpha \leq 1$, and let $L > 0$. A function $f : [0,1]^d \to \mathbb{R}$ called (β, L)-smooth if for every $(k_1, \ldots, k_d)$, $k_i \in \mathbb{N}_0$, and $\sum_{i=1}^d k_i = k$, the partial derivative $\partial^k f / (\partial x_1^{k_1} \cdots \partial x_d^{k_d})$ exists and satisfies

$$\left| \frac{\partial^k f}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}}(x) - \frac{\partial^k f}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}}(z) \right| \leq L \cdot \|x - z\|^\alpha \tag{4.1}$$

for all $x, z \in [0,1]^d$. The isotropic Hölder class, denoted $\mathcal{F}_1(\beta, L)$, is defined as the set of all (β, L)-smooth functions $f : [0,1]^d \to \mathbb{R}$.

Example 1. Suppose Assumptions 1–2 are satisfied. For any closed perturbation set $\Delta_n \in \mathbb{R}^d$, define

$$r_n \triangleq \max_{\delta_1, \delta_2 \in \Delta_n} \|\delta_1 - \delta_2\|. \tag{4.2}$$

If there exists a pair of δ and δ' in Δ_n such that $\|\delta - \delta'\| = r_n$ and $\{t\delta + (1-t)\delta' : 0 \leq t \leq 1\} \subseteq \Delta_n$, then we have

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_1(\beta, L)} R_{\Delta_n}(f, \hat{f}) \asymp \left(\frac{\log n}{n} \right)^{\frac{\beta}{2\beta + d}} + C_{d,\beta} r_n^{1 \wedge \beta}, \tag{4.3}$$

where $C_{d,\beta}$ is a constant depending on d and β, but independent of n.

In view of (3.10), the proof of the result in Example 1 consists of examining the standard minimax rate $\inf_{\hat{f}} \sup_{f \in \mathcal{F}_1(\beta,L)} R(f, \hat{f})$ and the rate of $\sup_{f \in \mathcal{F}_1(\beta,L)} L^*_{\Delta_n}(f)$. The standard minimax rate within the isotropic

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

Hölder class is established in Stone (1982), which demonstrates that

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_1(\beta, L)} R(f, \hat{f}) \asymp \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}}. \tag{4.4}$$

The determination of the rate of $\sup_{f \in \mathcal{F}_1(\beta,L)} L^*_{\Delta_n}(f)$ is provided in Section S1 of the Supplementary Material.

The quantity r_n in (4.2) measures the length of the longest line segment contained in the set Δ_n, and it may depend on the sample size n. The condition imposed on Δ_n is quite mild, which is satisfied by the ℓ_p-ball: $B_p^{q_n} \triangleq \{\delta \in \mathbb{R}^d : \|\delta\|_p \leq q_n\}$, $0 < p \leq \infty$, and the ℓ_p-ball with the ℓ_0-constraint: $B_p^{q_n} \cap \{\delta : \|\delta\|_0 \leq s_n\}$. Note that there is an extensive body of prior work studying adversarial machine learning based on ℓ_0 (Delgosha et al., 2024), ℓ_2 (Bhattacharjee and Chaudhuri, 2020; Bhattacharjee et al., 2021), and ℓ_∞ attacks (Athalye et al., 2018; Marzi et al., 2018). However, these analyses focus on the specific attacks and lack general applicability. In contrast, the result in Example 1 sheds theoretical insight on the adversarial robustness under the general ℓ_p-attacks with $0 < p \leq \infty$. Specifically, when $\Delta_n = B_p^{q_n}$, we have

$$2q_n \leq r_n = 2 \max_{\|\delta\|_p \leq q_n} \|\delta\|_2 = 2d^{\max\{0, \frac{1}{2} - \frac{1}{p}\}} q_n \leq 2d^{\frac{1}{2}} q_n,$$

and thus the minimax adversarial risk is given by

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_1(\beta, L)} R_{B_p^{q_n}}(f, \hat{f}) \asymp \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}} + C_{d,\beta} q_n^{1 \wedge \beta}, \tag{4.5}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

where $C_{d,\beta}$ is a constant depending on d and β, but independent of n. The optimal rate (4.5) can be reached by the adversarial plug-in estimator (3.4) with $\tilde{f}$ constructed by a suitably designed local polynomial estimator with optimal sup-norm convergence (see, e.g., Stone (1982), Gaïffas (2007), and Tsybakov (2008)).

The equation (4.5) shows that when $\beta < 1$ and $q_n \lesssim (\log n/n)^{1/(2\beta+d)}$, the minimax rate in the adversarial sup-norm remains unchanged to the standard minimax rate (4.4). However, as the magnitude of perturbation increases, e.g., $q_n \gtrsim (\log n/n)^{1/(2\beta+d)}$, the minimax risk has the order q_n^β. When $\beta \geq 1$ and the functions in $\mathcal{F}_1(\beta, L)$ become smoother, the critical radius q_n for the phase transition is $(\log n/n)^{\beta/(2\beta+d)}$. It is also worth noting that the norm parameter p, which controls the shape of the perturbation set $B_p^{q_n}$, does not affect the adversarial minimax rates in this example up to a constant that may depend on d and β. However, in other regression function classes of interest, the shape of the perturbation may have an effect on the robustness of a given estimator; see Section 4.2 for further discussion.

Remark 2. In this paper, we primarily focus on the adversarial sup-norm as the robustness performance measure. Using the uniformity of the sup-norm

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

loss, we can derive the following upper bound on the adversarial L_2-loss

$$\bar{L}_{\Delta_n}(f,\hat{f}) \triangleq \int_{\mathcal{X}} \sup_{\substack{\delta\in\Delta_n \\ x+\delta\in\mathcal{X}}} \left|f(x)-\hat{f}(x+\delta)\right|^2 \mathbb{P}_X dx \lesssim L^2_{\Delta_n}(f,\hat{f}),$$

under the assumption that $\mathcal{X}$ is a compact set and $\mathbb{P}_X$ satisfies Assumption 1. Based on this relation and (4.3), we can also derive an upper bound on the minimax adversarial risk under L_2-loss over the isotropic Hölder class:

$$\left(\frac{\log n}{n}\right)^{\frac{2\beta}{2\beta+d}} + C_{d,\beta} {r_n}^{2(1\wedge\beta)}.$$

It remains to be seen if this is the minimax optimal rate.

4.2 Anisotropic Hölder class

In practice, one of the typically desired properties of a regression function or its estimator is that it is invariant or robust against changes or perturbations of an input in some specific directions. For example, in image classification tasks, the target function should be invariant against a spatial shift or rotation of an input image (Simard et al., 2003; Krizhevsky et al., 2012). In the same spirit, in the context of autonomous driving, a traffic sign recognition model should be trained to be robust to natural variations in severe weather conditions.

Motivated by these examples, in this subsection, we investigate the adversarial minimax risks on the anisotropic Hölder class $\mathcal{F}_2(\beta, L)$, where

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

$\beta = (\beta_1, \ldots, \beta_d) \in (0,1]^d$ and $L = (L_1, \ldots, L_d) \in (0,\infty)^d$ (Birgé, 1986; Bertin, 2004a; Bhattacharya et al., 2014; Jeong and Rockova, 2023). This class is defined by

$$\mathcal{F}_2(\beta, L) \triangleq \big\{ f : [0,1]^d \to \mathbb{R} : |f(x) - f(z)| \leq L_1 |x_1 - z_1|^{\beta_1} + \cdots + L_d |x_d - z_d|^{\beta_d} \big\}, \tag{4.6}$$

which is a set of functions that have "direction-dependent" smoothness, whereas the isotropic Hölder class considered in Section 4.1 assumes isotropic smoothness that is uniform in all directions.

Example 2. Suppose Assumptions 1–2 hold. For any perturbation set $\Delta_n \in \mathbb{R}^d$, define $r_i \triangleq \sup_{\delta,\delta' \in \Delta_n} |\delta_i - \delta_i'|$ for $1 \leq i \leq d$, where $\delta = (\delta_1, \ldots, \delta_d)$ and $\delta' = (\delta_1', \ldots, \delta_d')$. Then we have

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_2(\beta, L)} R_{\Delta_n}(f, \hat{f}) \asymp \left(\frac{\log n}{n} \right)^{\frac{\bar{\beta}}{2\bar{\beta}+d}} + \max \left\{ r_1^{\beta_1}, \ldots, r_d^{\beta_d} \right\}, \tag{4.7}$$

where $\bar{\beta} = d / (\sum_{i=1}^d 1/\beta_i)$.

The first term on the right side of (4.7) represents the standard minimax rate under the sup-norm, which is determined by the average smoothness and the dimension d. The second term is related to the maximum deviation of function values along each coordinates. Combining the results in Section 3 with Bertin (2004a,b), it can be deduced that the adversarial minimax rate is achievable through the plug-in estimator (3.4), with $\tilde{f}$ being

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

a multivairate kernel estimator with different bandwidths across different coordinates.

To compare the adversarial minimax rates in the isotropic and anisotropic Hölder classes, let us consider a specific perturbation set $\Delta_n = \{\delta : |\delta_1| \leq q_n, \delta_2 = \cdots = \delta_d = 0\}$, where $q_n \to 0$ and $q_n \gtrsim (\log n/n)^{1/(2\bar{\beta}+d)}$. Note that the attacks within Δ_n are concentrated solely on the first coordinate. Suppose $\beta_1 > \bar{\beta}$. The isotropic Hölder class with the smoothness parameter $\bar{\beta}$ exhibits the minimax rate:

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_1(\bar{\beta}, L)} R_{\Delta_n}(f, \hat{f}) \asymp q_n^{\bar{\beta}}.$$

In contrast, for the anisotropic Hölder class, the minimax rate is:

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_2(\beta, L)} R_{\Delta_n}(f, \hat{f}) \asymp \max\{r_1^{\beta_1}, \ldots, r_d^{\beta_d}\} = q_n^{\beta_1},$$

which converges significantly faster than $\inf_{\hat{f}} \sup_{f \in \mathcal{F}_1(\bar{\beta}, L)} R_{\Delta_n}(f, \hat{f})$ as $q_n^{\beta_1}/q_n^{\bar{\beta}} \to 0$. This phenomenon implies that although the average smoothness is the same for the two function classes, when the attack is only in a smoother direction, the adversarial minimax risk in the anisotropic Hölder class is faster than that in the isotropic Hölder class.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

5. Simulation studies

In this section, we present several numerical experiments to illustrate the theoretical results established in Sections 3–4. The data are generated from the model (2.1), where $\mathcal{X} = [0,1]^2$, X follows a uniform distribution on $[0,1]^2$, and ξ is independent of X and distributed as $N(0,\sigma^2)$. We consider several regression functions and attack scenarios:

Case 1 $f(x_1,x_2) = \sqrt{x_1 x_2}$ with perturbation set $\Delta_n = B_\infty^r$.

Case 2 $f(x_1,x_2) = \sqrt{(x_1-0.5)^2+(x_2-0.5)^2}$, with $\Delta_n = B_\infty^r$.

Case 3 $f(x_1,x_2) = \sqrt{x_1}+0.1x_2-0.5$, with $\Delta_n = [-4r,+4r]\times[-r/4,+r/4]$.

Case 4 $f(x_1,x_2) = \sqrt{x_1}+0.1x_2-0.5$, with $\Delta_n = [-r/4,+r/4]\times[-4r,+4r]$.

In each case, σ^2 is adjusted so that the signal-to-noise ratio equals 5. The attack magnitude r increases from 0 to 0.1. Cases 1–2 serve as two representative examples of isotropic Hölder classes, where the perturbation set is chosen as the ℓ_∞-ball. In contrast, Cases 3–4 consider regression functions with different degrees of variation along different axes, where the attack magnitudes are also anisotropic.

We consider three competing methods. The baseline method (LP) is the classical local polynomial regression studied in Stone (1982), Bertin

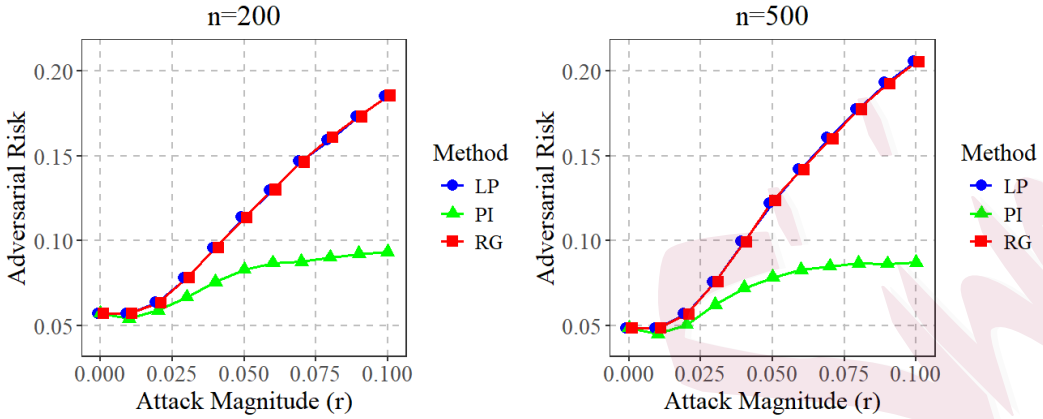
Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

(2004b), and Gaïffas (2007) based on the rectangular kernel. We employ a polynomial of degree $\ell = 1$ (i.e., local linear regression). In Cases 1 and 2, the bandwidth is set as $h = n^{-1/(0.5+2)}$ and $h = n^{-1/(1+2)}$, respectively. In Cases 3 and 4, we use different bandwidths for different coordinates, setting $h_1 = n^{-1/(0.5+2)}$ and $h_2 = n^{-1/(1+2)}$. These choices are theoretically proven to achieve the standard minimax rates in the respective cases.

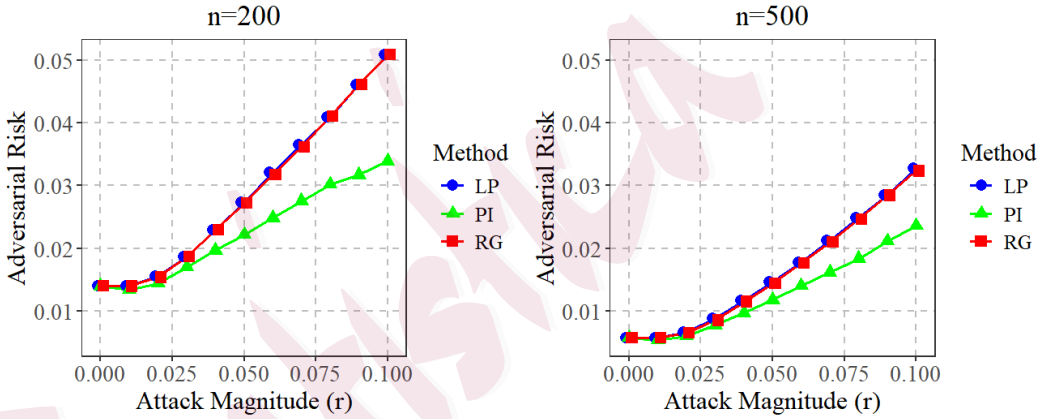
Building on the LP method, we consider two additional competing methods. The first (PI) follows (3.4), where $\tilde{f}$ is the LP estimator. The second method is a ridge-type local polynomial estimator (RG), which follows the LP approach but incorporates a ridge penalty with parameter r^2 on the linear coefficients during the estimation of the LP coefficients. The ridge-type strategy can be seen as an approximation of adversarial training (Ribeiro and Schön, 2023) and has also been proven to possess desirable robustness properties under several specific setups (Zhang et al., 2019; Xing et al., 2021). Figures 1–2 present the adversarial risk for the three competing methods over 100 simulation replications. In each replication, the adversarial loss is evaluated at 100 uniformly sampled points in $[0, 1]^2$.

From Figures 1–2, we observe a significant advantage of the PI method over the classical LP method and its ridge-type variant. For instance, in Case 2 with $n = 200$ and $r = 0.5$, the adversarial risk and its standard

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298



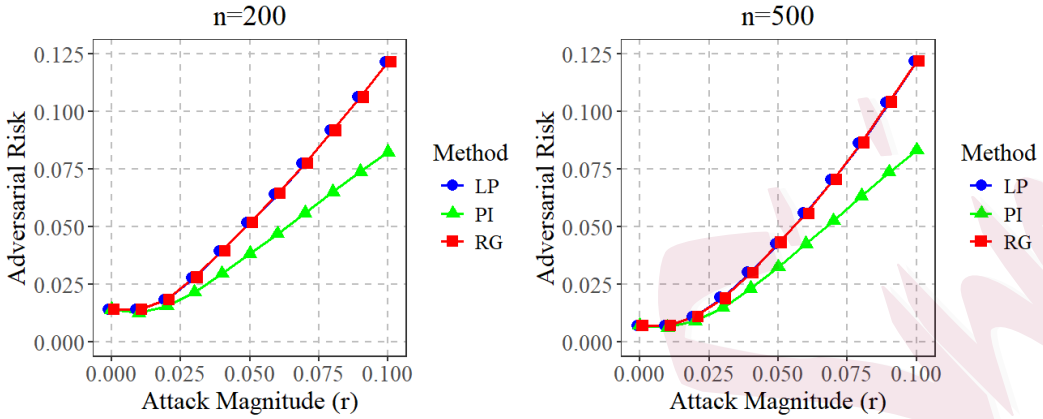
(a) Case 1



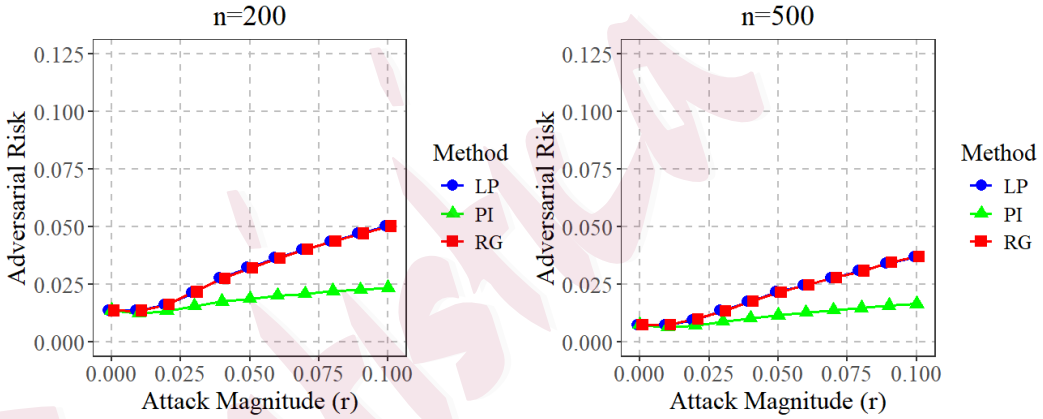
(b) Case 2

Figure 1: Adversarial risk for the three competing methods as the attack magnitude increases: panel (a) corresponds to Case 1, and panel (b) corresponds to Case 2.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298



(a) Case 3



(b) Case 4

Figure 2: Adversarial risk for the three competing methods as the attack magnitude increases: panel (a) corresponds to Case 3, and panel (b) corresponds to Case 4.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

error for LP, RG, and PI are 2.73e-2 (0.001), 2.71e-2 (0.001), and 2.21e-2 (0.001), respectively. These results demonstrate that the adversarial plug-in procedure 3.4 achieves a substantial improvement in robustness compared to the other two methods. The patterns depicted in Figures 1–2 further corroborate the insights discussed in Sections 4.1–4.2. For example, in Case 1, where the regression function belongs to $\mathcal{F}_1(\beta, L)$ with $\beta = 1/2$, the adversarial risk curve exhibits a concave shape, consistent with the $r^{1/2}$. In Case 2, the adversarial risk curve is approximately linear as r increases, which aligns with Example 1 that the adversarial loss in this case is dominated by r when r is large. Additionally, Figure 2 reveals that strong attacks along directions with higher variability can significantly degrade the performance of competing methods, supporting the theoretical results presented in Example 2.

Furthermore, although existing literature suggests that ridge-type regularization can enhance adversarial robustness under various modeling frameworks (see, e.g., Zhang et al., 2019; Xing et al., 2021), its effectiveness in the context of local nonparametric estimation remains limited. This limitation arises because ridge regularization in RG primarily controls the variation of the LP estimator at a given local point but does not regulate the variation of the estimator across different local points. Consequently, the RG method

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

may still be vulnerable to adversarial attacks under our context.

6. Discussion

In this paper, we focus on the nonparametric regression problem under the adversarial attacks and examine the minimax rates of convergence in the adversarial sup-norm. Unlike the minimax analysis for the specific models in Dan et al. (2020) and Xing et al. (2021), the results established in this paper are of a general nature. They are applicable across diverse regression function classes and arbitrary additive perturbation sets. We show that the minimax rate in the adversarial setting exhibits a modular form, which equals the standard minimax rate in the absence of an adversary, plus the maximum deviation of the true function value within the perturbation set. Applying the general results to specific models is straightforward: it entails determining the standard minimax rate and calculating the largest Lipschitz constant of the functions in the target class. We further investigate two nonparametric function classes, illuminating the impacts of the different perturbation sets on the adversarial minimax rates.

It should be pointed out that the proposed adversarial plug-in estimation procedure in this paper is nonadaptive, since it depends on information about the unknown perturbation set Δ_n. In the context of practical appli-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

cations, an important direction for future research is to develop estimation procedures that are both adaptive across different function classes and unknown perturbation sets. Another direction is deriving the minimax rates in the general L_p-norm under the adversarial attacks. In the standard setting, it is well-known that the metric entropy of the regression function class plays a fundamental role in determining the minimax rates of convergence (LeCam, 1973; Birgé, 1986; Yatracos, 1985; Yang and Barron, 1999). Extending these general theories to the adversarial setting is of great interest.

Supplementary Material

The online Supplementary Material includes the theoretical proofs for the results in the main theorems and Examples 1-2.

Acknowledgments

The authors thank the associate editor and two anonymous referees for their insightful and constructive comments.

References

Athalye, A., N. Carlini, and D. Wagner (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International*

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

Conference on Machine Learning, Volume 80, pp. 274–283.

Attias, I., A. Kontorovich, and Y. Mansour (2019). Improved generalization bounds for robust learning. In *Algorithmic Learning Theory*, Volume 98, pp. 162–183.

Audibert, J.-Y. and A. B. Tsybakov (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics 35*(2), 608–633.

Awasthi, P., N. Frank, and M. Mohri (2020). Adversarial learning guarantees for linear hypotheses and neural networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, Volume 119, pp. 431–441.

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

Bertin, K. (2004a). Asymptotically exact minimax estimation in sup-norm for anisotropic Hölder classes. *Bernoulli 10*(5), 873–888.

Bertin, K. (2004b). Minimax exact constant in sup-norm for nonparametric regression with random design. *Journal of Statistical Planning and Inference 123*(2), 225–242.

Bhagoji, A. N., D. Cullina, and P. Mittal (2019). Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems 32*, pp. 7496–7508.

Bhattacharjee, R. and K. Chaudhuri (2020). When are non-parametric methods robust? In *International Conference on Machine Learning*, pp. 832–841. PMLR.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

PMLR.

Bhattacharya, A., D. Pati, and D. Dunson (2014). Anisotropic function estimation using multi-bandwidth Gaussian processes. *The Annals of Statistics 42*(1), 352–381.

Biggio, B., I. Corona, D. Maiorca, B. Nelson, N. Srndic, P. Laskov, G. Giacinto, and F. Roli (2013). Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases - European Conference*, Volume 8190 of *Lecture Notes in Computer Science*, pp. 387–402. Springer.

Birgé, L. (1986). On estimating a density using hellinger distance and some other strange facts. *Probability theory and related fields 71*(2), 271–291.

Carlini, N. and D. Wagner (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57.

Chen, X. and T. M. Christensen (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics 188*(2), 447–465.

Cohen, J., E. Rosenfeld, and Z. Kolter (2019). Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, Volume 97, pp. 1310–1320.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research 12*(76), 2493–2537.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

REFERENCES

Dan, C., Y. Wei, and P. Ravikumar (2020). Sharp statistical guaratees for adversarially robust Gaussian classification. In *Proceedings of the 37th International Conference on Machine Learning*, Volume 119, pp. 2345–2355.

Delgosha, P., H. Hassani, and R. Pedarsani (2024). Binary classification under ℓ_0 attacks for general noise distribution. *IEEE Transactions on Information Theory 70*(2), 1284–1299.

Devroye, L. (1978). The uniform convergence of nearest neighbor regression function estimators and their application in optimization. *IEEE Transactions on Information Theory 24*(2), 142–151.

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

Donoho, D. L. (1994). Asymptotic minimax risk for sup-norm loss: Solution via optimal recovery. *Probability Theory and Related Fields 99*, 145–170.

Duchi, J., T. Hashimoto, and H. Namkoong (2023). Distributionally robust losses for latent covariate mixtures. *Operations Research 71*(2), 649–664.

Finlay, C. and A. M. Oberman (2021). Scaleable input gradient regularization for adversarial robustness. *Machine Learning with Applications 3*, 100017.

Gaïffas, S. (2007). Sharp estimation in sup norm with random design. *Statistics & Probability Letters 77*(8), 782–794.

Giné, E. and R. Nickl (2009). Uniform limit theorems for wavelet density estimators. *The Annals of Probability 37*(4), 1605–1646.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

REFERENCES

Goodfellow, I. J., J. Shlens, and C. Szegedy (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Grigorescu, S., B. Trasnea, T. Cocias, and G. Macesanu (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics 37*(3), 362–386.

Ibragimov, I. and R. Khas' minskii (1982). Bounds for the risks of non-parametric regression estimates. *Theory of Probability & Its Applications 27*(1), 84–99.

Imaizumi, M. (2023). Sup-norm convergence of deep neural network estimator for nonparametric regression by adversarial training. *arXiv preprint arXiv:2307.04042*.

Javanmard, A. and M. Soltanolkotabi (2022). Precise statistical analysis of classification accuracies for adversarial training. *The Annals of Statistics 50*(4), 2127–2156.

Javanmard, A., M. Soltanolkotabi, and H. Hassani (2020). Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, Volume 125, pp. 2034–2078.

Jeong, S. and V. Rockova (2023). The art of bart: Minimax optimality over nonhomogeneous smoothness in high dimension. *Journal of Machine Learning Research 24*(337), 1–65.

Khim, J. and P.-L. Loh (2018). Adversarial risk bounds via function transformation. *arXiv preprint arXiv:1810.09519*.

Korostelev, A. and M. Nussbaum (1999). The asymptotic minimax constant for sup-norm loss in nonparametric density estimation. *Bernoulli 5*(6), 1099–1118.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, Volume 25,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

pp. 1106–1114.

LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics 1*(1), 38–53.

Lee, J. and M. Raginsky (2018). Minimax statistical learning with wasserstein distances. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 2692–2701.

Lepski, O. V. and A. B. Tsybakov (2000). Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point. *Probability Theory and Related Fields 117*(1), 17–48.

Liu, C., Y. Jiao, J. Wang, and J. Huang (2023). Non-asymptotic bounds for adversarial excess risk under misspecified models. *arXiv preprint arXiv:2309.00771*.

Madry, A., A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Marzi, Z., S. Gopalakrishnan, U. Madhow, and R. Pedarsani (2018). Sparsity-based defense against adversarial attacks on linear classifiers. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 31–35. IEEE.

Mehrabi, M., A. Javanmard, R. A. Rossi, A. Rao, and T. Mai (2021). Fundamental tradeoffs in distributionally adversarial training. In *Proceedings of the 38th International Conference on Machine Learning*, Volume 139, pp. 7544–7554.

Min, Y., L. Chen, and A. Karbasi (2021). The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. In *Proceedings of the Thirty-Seventh*

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

Conference on Uncertainty in Artificial Intelligence, Volume 161, pp. 129–139.

Montasser, O., S. Hanneke, and N. Srebro (2019). Vc classes are adversarially robustly learnable, but only improperly. In *Proceedings of the Thirty-Second Conference on Learning Theory*, Volume 99, pp. 2512–2530.

Moosavi-Dezfooli, S., A. Fawzi, and P. Frossard (2016). Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 2574–2582.

Papernot, N., P. McDaniel, S. Jha, M. Fredrikson, Z. Celik, and A. Swami (2016). The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 372–387.

Pydi, M. S. and V. S. Jog (2020). Adversarial risk via optimal transport and optimal couplings. In *Proceedings of the 37th International Conference on Machine Learning*, Volume 119, pp. 7814–7823.

Raghunathan, A., J. Steinhardt, and P. Liang (2018). Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*.

Raghunathan, A., S. M. Xie, F. Yang, J. C. Duchi, and P. Liang (2019). Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*.

Ribeiro, A. H. and T. B. Schön (2023). Overparameterized linear regression under adversarial attacks. *IEEE Transactions on Signal Processing 71*, 601–614.

Robey, A., H. Hassani, and G. J. Pappas (2020). Model-based robust deep learning: Generalizing

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

to natural, out-of-distribution data. *arXiv preprint arXiv:2005.10247*.

Schmidt, L., S. Santurkar, D. Tsipras, K. Talwar, and A. Madry (2018). Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems 31*, pp. 5019–5031.

Schmidt-Hieber, J. and P. Zamolodtchikov (2024). Local convergence rates of the nonparametric least squares estimator with applications to transfer learning. *Bernoulli 30*(3), 1845–1877.

Shapiro, A., D. Dentcheva, and A. Ruszczynski (2021). *Lectures on Stochastic Programming: Modeling and Theory*. SIAM.

Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis (2016). Mastering the game of Go with deep neural networks and tree search. *Nature 529*(7587), 484–489.

Simard, P., D. Steinkraus, and J. Platt (2003). Best practices for convolutional neural networks applied to visual document analysis. In *Seventh International Conference on Document Analysis and Recognition*, pp. 958–963.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics 10*(4), 1040–1053.

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

Tsipras, D., S. Santurkar, L. Engstrom, A. Turner, and A. Madry (2019). Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations*.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer New York, NY.

Tu, Z., J. Zhang, and D. Tao (2019). Theoretical analysis of adversarial learning: A minimax approach. In *Advances in Neural Information Processing Systems*, Volume 32, pp. 12259–12269.

Wang, Y., S. Jha, and K. Chaudhuri (2018). Analyzing the robustness of nearest neighbors to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, Volume 80, pp. 5133–5142.

Xing, Y., Q. Song, and G. Cheng (2021). Predictive power of nearest neighbors algorithm under random perturbation. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, Volume 130, pp. 496–504.

Xing, Y., R. Zhang, and G. Cheng (2021). Adversarially robust estimate and risk analysis in linear regression. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, Volume 130, pp. 514–522.

Yang, Y. and A. Barron (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics 27*(5), 1564–1599.

Yang, Y.-Y., C. Rashtchian, Y. Wang, and K. Chaudhuri (2020). Robustness for non-parametric classification: a generic attack and defense. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Volume 108, pp. 941–951.

Yatracos, Y. G. (1985). Rates of convergence of minimum distance estimators and kolmogorov's

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0298

entropy. *The Annals of Statistics 13*(2), 768–774.

Yin, D., K. Ramchandran, and P. L. Bartlett (2019). Rademacher complexity for adversarially robust generalization. In *Proceedings of the 36th International Conference on Machine Learning*, Volume 97, pp. 7085–7094.

Zhang, H., Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan (2019). Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning*, Volume 97, pp. 7472–7482.

Zhao, P. and Z. Wan (2024). Robust nonparametric regression under poisoning attack. *Proceedings of the AAAI Conference on Artificial Intelligence 38*(15), 17007–17015.

Jingfu Peng

Yau Mathematical Sciences Center, Tsinghua University

E-mail: aypjf@mail.tsinghua.edu.cn

Yuhong Yang

Yau Mathematical Sciences Center, Tsinghua University Beijing Institute of Mathematical Sciences and Applications

E-mail: yyangsc@mail.tsinghua.edu.cn