Nonasymptotic Bounds for Adversarial Excess Risk under Misspecified Models*

Changyu Liu[†], Yuling Jiao[‡], Junhui Wang[†], and Jian Huang[§]

Abstract. We propose a general approach to evaluating the performance of robust estimators based on adversarial losses under misspecified models. We first show that adversarial risk is equivalent to the risk induced by a distributional adversarial attack under certain smoothness conditions. This ensures that the adversarial training procedure is well-defined. To evaluate the generalization performance of the adversarial estimator, we study the adversarial excess risk. Our proposed analysis method includes investigations on both generalization error and approximation error. We then establish nonasymptotic upper bounds for the adversarial excess risk associated with Lipschitz loss functions. In addition, we apply our general results to adversarial training for classification and regression problems. For the quadratic loss in nonparametric regression, we show that the adversarial excess risk bound can be improved over that for a general loss.

Key words. adversarial attack, approximation error, generalization, misspecified model, robustness

MSC codes. 62G05, 62G35, 68T07

DOI. 10.1137/23M1598210

1. Introduction. Deep learning methods are known to be vulnerable to adversarial examples, which are formed by applying an imperceptible perturbation to the input such that the perturbed input causes the model to make a highly confident but erroneous prediction [43, 17]. The problem gained widespread attention in recent years. Methods for finding adversarial attacks [17, 32, 30, 11, 8, 1, 42] and developing adversarial defense [33, 28, 53, 13] have been extensively studied. Among the adversarial defense methods, adversarial training [28] has been empirically proven to be successful.

Although there has been significant progress in developing methods for defending adversarial attacks, theoretical understanding of adversarial robustness remains limited. [34, 35] considered the classification loss under the adversarial binary classification setting and obtained

^{*}Received by the editors September 1, 2023; accepted for publication (in revised form) April 22, 2024; published electronically October 1, 2024. Changyu Liu and Yuling Jiao contributed equally to this work.

https://doi.org/10.1137/23M1598210

Funding: The work of the second author was supported by the National Nature Science Foundation of China (grant 12371441), by the Fundamental Research Funds for the Central Universities, and by the research fund of KLATASDSMOE of China. The work of the third author was supported in part by HK RGC GRF-14306523 and CUHK Startup Grant 4937091. The work of the fourth author was supported by the National Natural Science Foundation of China (grant 72331005) and research grants from The Hong Kong Polytechnic University.

[†]Department of Statistics, The Chinese University of Hong Kong, Hong Kong SAR, China (changyuliu@ cuhk.edu.hk, junhuiwang@cuhk.edu.hk).

[‡]School of Mathematics and Statistics, and Hubei Key Laboratory of Computational Science, Wuhan University, Wuhan 430072, China (yulingjiaomath@whu.edu.cn).

[§]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong SAR, China (j.huang@polyu.edu.hk).

Comparison of recent methods for studying generalization performance of an adversarial estimator. The ℓ_r attack refers to general ℓ_r adversarial attacks for $r \geq 1$, the error \mathcal{E}_{gen} refers to the generalization error, and \mathcal{E}_{app} refers to the approximation error.

Table 1

	ℓ_r attack	FNNs	\mathcal{E}_{gen}	\mathcal{E}_{app}
[52]	×	×	\checkmark	X
[23]	\checkmark	\checkmark	\checkmark	X
[3]	\checkmark	×	\checkmark	X
[31]	\checkmark	\checkmark	\checkmark	X
[46]	\checkmark	\checkmark	\checkmark	X
This paper	\checkmark	\checkmark	\checkmark	\checkmark

the optimal adversarial classification risk. [35, 4] and [10] proved the existence and minimax properties for the adversarial classification risk. The results were later extended by [16] to the setting where surrogate functions were used. Another series of work investigated the calibration and consistency of the surrogate loss functions under adversarial attacks [6, 2, 5, 29].

Several authors have considered the generalization errors of adversarial estimators in recent years. Examples include [52, 23, 3, 31], who analyzed the Rademacher complexity of adversarial loss function class. [46] transformed the adversarial learning into a distributional robustness optimization (DRO) problem and studied its generalization properties. However, the above work only considered well-specified models, i.e., the underlying target function is assumed to belong to a class of neural network functions. As is well known, in the classical nonparametric method for classification and regression, the underlying target is the link function defined as $\mathbb{E}[Y|X = x]$, which is not a neural network function in general. Therefore, a natural question would be, what are the properties of an adversarial estimator under misspecified models, i.e., when the underlying target function is not an exact neural network function, but can only be approximated by neural networks? Under this more general setting, it is necessary to consider both the generalization error and the approximation error caused by model misspecification. In this paper, we study this problem systematically. We first provide a summary of the main features of our result and the related ones in Table 1 below.

While adversarial training improves the robustness of an estimator on adversarially perturbed data, this benefit often comes at the cost of more resource consumption and leads to a reduction of accuracy on natural unperturbed data [44]. Some recent works have tried to gain theoretical understanding of the trade-offs between accuracy and robustness [28, 38, 36, 44, 37, 15, 27, 53, 20, 19]. However, none of the above-mentioned works studied the setting of deep adversarial training with misspecified models.

In this work, we provide theoretical guarantees for deep adversarial training with misspecified models by establishing nonasymptotic error bounds for the adversarial excess risk, defined as the difference between the adversarial risk of an adversarial estimator and the optimal adversarial risk. The adversarial excess risk can be decomposed as

Adversarial excess risk $\leq \mathcal{E}_{gen} + \mathcal{E}_{app}$,

Table 2

Summary of error bounds (up to a logarithmic factor) in the paper, where ε represents the adversarial attack level, $r_1 = \alpha/(2d+3\alpha)$, $r_2 = 2\alpha/(2d+5\alpha)$, $r_3 = (d+3\alpha-1)/(2d+3\alpha)$, $r_4 = (d+1)/(2d+3\alpha)$, and $r_5 = (d+1)/(2d+5\alpha)$.

Loss function	Measurement	Error bound
Lipschitz	adversarial excess risk	$n^{-r_1} + n^{-r_3}\varepsilon$
	excess risk	$n^{-r_1} + \varepsilon$
	local worst-case excess risk	$n^{-r_1} + n^{r_4}\varepsilon$
Classification	adversarial excess risk	$n^{-r_1} + \varepsilon$
Quadratic	L ₂ -norm	$n^{-r_2} + n^{r_5}\varepsilon$

where \mathcal{E}_{gen} represents the generalization error and \mathcal{E}_{app} represents the approximation error. The explicit expressions of \mathcal{E}_{gen} and \mathcal{E}_{app} are given in Theorem 3.1. The adversarial setting poses significant challenges, particularly in analyzing supremum-type loss functions. These functions may lack measurability, even when considering measurable loss and estimation functions, and their analysis introduces complexities in assessing generalization and approximation errors. To address these challenges, we leverage the Lipschitz property of the loss and estimation functions and derive explicitly upper bounds for both generalization error and approximation error. These bounds reflect how the neural network's structure and the adversarial attack level influence the adversarial excess risk. Additionally, in cases where Lipschitzness of the loss function cannot be guaranteed, as for the quadratic loss function discussed in section 4.2, we provide a nonasymptotic error bound for the expectation of the adversarial excess risk.

Our main contributions are summarized as follows:

- We establish nonasymptotic error bounds for the adversarial excess risks under misspecified models and use the feedforward neural networks (FNNs) with constraints on the Lipschitz property. The error bounds explicitly illustrate the influence of the adversarial attack level and can achieve the rate $O(n^{-\alpha/(2d+3\alpha)})$ up to a logarithmic factor, where α represents the smoothness level of the underlying target function and d is the dimension of input. The structure of the neural network is specified to show when the error rate can be achieved.
- We also evaluate the adversarial estimator under natural risk and local worst-case risk.
- We apply our general results to the classification and nonparametric regression problems in an adversarial setting and establish nonasymptotic error bounds for the adversarial estimators under the adversarial classification risk and L_2 -norm, respectively.

The results for error bounds for the adversarial estimator in different settings and measurements are summarized in Table 2.

The rest of the paper is organized as follows. Section 2 introduces the notation and problem setup. Section 3 contains the main results of the paper. Section 4 presents applications of the results to classification and regression problems. In section 5, discussion on some related works is given. Concluding remarks are given in section 6. The proof of the main theorem is given in the appendix, and the remaining proofs are relegated to the supplementary material (supplement.pdf [local/web 358KB]).

Notation. Let the set of positive integers be denoted by $\mathbb{N} = \{1, 2, ...\}$ and let $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. If a and b are two quantities, we use $a \leq b$ or $b \geq a$ to denote the statement that $a \leq Cb$ for some constant C > 0. We denote $a \approx b$ when $a \leq b \leq a$. Let $\lceil a \rceil$ denote the smallest integer larger than or equal to quantity a. For a vector \boldsymbol{x} and $p \in [1, \infty]$, we use $\|\boldsymbol{x}\|_p$ to denote the p-norm of \boldsymbol{x} . For a function f, we use $\|f\|_{\infty}$ to denote the supremum norm of f.

2. Problem setup. In this section, we introduce the definition of adversarial risk and present a basic setup of adversarial training. We also lay the foundation for the theoretical analysis of adversarial training and establish some basic properties of adversarial risk.

2.1. Adversarial risk. Suppose that (X, Y) follows an unknown distribution P over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$. For a loss function $\ell : \mathbb{R} \times \mathcal{Y} \mapsto [0, \infty)$ and a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, the (population) natural risk is defined by

$$\mathcal{R}_P(f) = \mathbb{E}_{(X,Y)\sim P}[\ell(f(X),Y)].$$

To evaluate the performance of function f in the presence of adversarial attacks, the (population) adversarial risk is defined by

$$\widetilde{\mathcal{R}}_P(f,\varepsilon) = \mathbb{E}_{(X,Y)\sim P} \left[\sup_{X' \in B_{\varepsilon}(X)} \ell(f(X'), Y) \right],$$

where $B_{\varepsilon}(\boldsymbol{x}) = \{\boldsymbol{x}' \in \mathcal{X} : \|\boldsymbol{x}' - \boldsymbol{x}\|_{\infty} \leq \varepsilon\}$. Here we focus on ℓ_{∞} attack. In section 3, we will show that the proposed analysis method can be easily extended to a general ℓ_r attack.

2.2. Properties of adversarial risk. Adversarial risk has been widely considered in recent years for the goal of deriving adversarial robust estimators. To facilitate the analysis, we make the following assumptions.

Assumption 2.1. $\mathcal{Z} \subseteq [0,1]^d \times [-1,1]$ and $\cup_{\boldsymbol{x} \in \mathcal{X}} B_{\varepsilon}(\boldsymbol{x}) \subseteq [0,1]^d$ hold for $\varepsilon > 0$.

The assumption $\cup_{\boldsymbol{x}\in\mathcal{X}}B_{\varepsilon}(\boldsymbol{x})\subseteq[0,1]^d$ is to guarantee that the estimation function class is well-defined under the adversarial setting. Our analysis can be easily extended to a more general setting, where \mathcal{Z} is bounded. For a loss function $\ell:\mathbb{R}\times\mathcal{Y}\mapsto[0,\infty)$, we define

$$\operatorname{Lip}^{1}(\ell) = \sup_{y \in \mathcal{Y}} \sup_{u_{1} \neq u_{2}} \frac{|\ell(u_{1}, y) - \ell(u_{2}, y)|}{|u_{1} - u_{2}|}$$

Assumption 2.2. The loss function is continuous and satisfies $\operatorname{Lip}^1(\ell) < \infty$.

The assumption $\operatorname{Lip}^1(\ell) < \infty$ is weaker than the Lipschitz continuity condition, since it only imposes restriction on the Lipschitz constant of $\ell(\cdot, y)$ for every $y \in \mathcal{Y}$. The assumption is satisfied by many commonly used loss functions, such as the hinge loss and ρ -margin loss.

We first show that the adversarial risk is well-defined in our setting. This is necessary since the adversarial risk may not be well-defined in general [35]. Specifically, we show that the adversarial risk for a function f can be represented by a natural risk with the expectation taken over a shifted distribution. Moreover, the distance between the shifted distribution and the data generating distribution can be measured by the ∞ th Wasserstein distance. Let $d_{\mathcal{Z}}$ denote a metric over \mathcal{Z} satisfying $d_{\mathcal{Z}}(\boldsymbol{z}_1, \boldsymbol{z}_2) = \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_{\infty} + |y_1 - y_2|$ for any $\boldsymbol{z}_1 = (\boldsymbol{x}_1, y_1)$ and $\boldsymbol{z}_2 = (\boldsymbol{x}_2, y_2) \in \mathcal{Z}$. Let $\mathcal{P}(\mathcal{Z})$ denote the space of Borel probability measures on \mathcal{Z} . For $p \in [1, \infty)$, the *p*th Wasserstein distance between two probability measures $P, Q \in \mathcal{P}(\mathcal{Z})$ is defined as

$$W_p(P,Q) = \left\{ \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_{(Z_1,Z_2) \sim \pi} [d_{\mathcal{Z}}(Z_1,Z_2)^p] \right\}^{1/p},$$

where $\Pi(P,Q)$ denotes the collection of all probability measures on $\mathcal{Z} \times \mathcal{Z}$ with marginals P and Q. The ∞ th Wasserstein distance is defined to be the limit of the pth Wasserstein distances, which can also be characterized by

$$W_{\infty}(P,Q) = \inf_{\pi \in \Pi(P,Q)} \operatorname{ess\,sup}_{(Z_1,Z_2) \sim \pi} d_{\mathcal{Z}}(Z_1,Z_2).$$

Since $W_p(P,Q) \leq W_q(P,Q)$ for any $1 \leq p \leq q \leq \infty$, the ∞ th Wasserstein distance is stronger than any *p*th Wasserstein distance. Similarly, for $p \in [1,\infty]$, we define the *p*th Wasserstein distance over $\mathcal{P}(\mathcal{X})$ based on the supremum norm, where $\mathcal{P}(\mathcal{X})$ denotes the space of Borel probability measures on \mathcal{X} .

Lemma 2.3. Suppose Assumption 2.1 holds, ℓ and f are continuous, and there exists a measurable function T^* satisfying $T^*(\mathbf{z}) \in B_{\varepsilon}(\mathbf{x})$ such that $\sup_{\mathbf{x}' \in B_{\varepsilon}(\mathbf{x})} \ell(f(\mathbf{x}'), y) = \ell(f(T^*(\mathbf{z})), y)$. Let the joint distribution of $(T^*(Z), Y)$ be denoted by P^* ; we have $W_{\infty}(P^*, P) \leq \varepsilon$ and

$$\mathcal{R}_P(f,\varepsilon) = \mathcal{R}_{P^\star}(f).$$

Lemma 2.3 shows that the adversarial risk is well-defined with respect to the shifted distribution P^* . With further analysis of this shifted distribution, we construct an equivalent relationship between the adversarial risk and the risk induced by the distribution-perturbing adversary [34, 35]. The result is given in section 2.3 below.

To state the next lemma, we denote the Lipschitz constant for a function f by

$$\operatorname{Lip}(f) = \sup_{\boldsymbol{x}_1 \neq \boldsymbol{x}_2} \frac{|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)|}{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_{\infty}}.$$

Lemma 2.4. Suppose Assumptions 2.1–2.2 hold and $\operatorname{Lip}(f) < \infty$. Then

$$\mathcal{R}_P(f) \leq \mathcal{R}_P(f,\varepsilon) \leq \mathcal{R}_P(f) + \operatorname{Lip}^1(\ell)\operatorname{Lip}(f)\varepsilon.$$

Lemma 2.4 shows that adversarial robustness is closely related to the Lipschitz constraints. This connection between robustness and Lipschitz constraints has been the subject of several studies. For instance, [9] identified a relationship between these two aspects in scenarios where the data distribution satisfies isoperimetric conditions. Studies such as [12] and [45] have utilized Lipschitz constraints to devise novel regularization techniques and training strategies for deep neural networks. While their empirical findings have underscored the effectiveness and robustness of these methodologies across well-established image classification datasets, a theoretical analysis of adversarial estimators was not within their scope. In contrast, our research is framed within a more general context, without specific distributional assumptions or task-related limitations, thereby broadening its relevance to a variety of problem settings. Moreover, our emphasis is on the theoretical examination of the estimator's properties, aiming to furnish a more comprehensive understanding of its behavior.

2.3. Relationship between adversarial risk and distribution-perturbing risk. We further study the shifted distribution in Lemma 2.3 and construct a relationship between the adversarial risk and another kind of risk induced by a distributional adversarial attack defined below.

For any distribution $Q \in \mathcal{P}(\mathcal{Z})$, we denote its corresponding pair of variables by (X, Y), and let the conditional distribution of \widetilde{X} given $\widetilde{Y} = y$ be denoted by Q_y for every $y \in \mathcal{Y}$. The collection of distributions Γ_{ε} is defined by

$$\Gamma_{\varepsilon} = \Big\{ Q \in \mathcal{P}(\mathcal{Z}) : \text{when } (\widetilde{X}, \widetilde{Y}) \sim Q, \text{ then } \widetilde{Y} \sim P_Y \text{ and } W_{\infty}(Q_y, P_y) \leq \varepsilon \, \forall y \in \mathcal{Y} \Big\},\$$

where P_y is the conditional distribution of X given Y = y and P_Y is the distribution of Y. Intuitively, it is helpful to think of Γ_{ε} as a collection of distributional adversarial attacks. For every $Q \in \Gamma_{\varepsilon}$, by observing sample (X, Y), with y denoting the value of Y, it perturbs X to \widetilde{X} such that $\widetilde{X} \sim Q_y$ and lets Q_y lie in an uncertainty set around P_y . This kind of adversarial attack strategy is also known as a distribution-perturbing adversary [34, 35]. The corresponding distribution-perturbing risk is defined as $\sup_{Q \in \Gamma_{\varepsilon}} \mathcal{R}_Q(f)$.

Theorem 2.5. Suppose Assumption 2.1 holds, and ℓ and f are continuous. Then we have $\widetilde{\mathcal{R}}_P(f,\varepsilon) = \sup_{Q \in \Gamma_{\varepsilon}} \mathcal{R}_Q(f).$

This theorem shows that the risks induced by the two different types of adversarial attack are equivalent under the smoothness condition. Similar results were constructed in [35, 34], where [35] focused on the binary classification setting and [34] considered the case when \mathcal{Y} is a discrete set of labels. From the proof of Theorem 2.5, we also show that $P^* \in \Gamma_{\varepsilon}$, which directly implies $W_{\infty}(P^*, P) \leq \varepsilon$. Hence, this shows a stronger relationship with P.

2.4. Adversarial training. Adversarial training aims to learn a target function f^* that minimizes the adversarial risk. In this work, we assume that f^* belongs to a Hölder class.

Definition 2.6 (Hölder class). Let $d \in \mathbb{N}$ and $\alpha = r + \beta > 0$, where $r \in \mathbb{N}_0$ and $\beta \in (0, 1]$. Let $s \in \mathbb{N}_0^d$ denote the multi-index. The Hölder class $\mathcal{H}^{\alpha}(\mathbb{R}^d)$ is defined as

$$\begin{aligned} \mathcal{H}^{\alpha}(\mathbb{R}^{d}) = & \left\{ f: \mathbb{R}^{d} \to \mathbb{R}, \max_{\|\boldsymbol{s}\|_{1} \leq r} \sup_{\boldsymbol{x} \in \mathbb{R}^{d}} |\partial^{\boldsymbol{s}} f(\boldsymbol{x})| \leq 1, \\ & \max_{\|\boldsymbol{s}\|_{1} = r} \sup_{\boldsymbol{x}_{1} \neq \boldsymbol{x}_{2}} \frac{|\partial^{\boldsymbol{s}} f(\boldsymbol{x}_{1}) - \partial^{\boldsymbol{s}} f(\boldsymbol{x}_{2})|}{\|\boldsymbol{x}_{1} - \boldsymbol{x}_{2}\|_{\infty}^{\beta}} \leq 1 \right\}. \end{aligned}$$

We let $\mathcal{H}^{\alpha} = \{ f : [0,1]^d \to \mathbb{R}, f \in \mathcal{H}^{\alpha}(\mathbb{R}^d) \}$ denote the restriction of $\mathcal{H}^{\alpha}(\mathbb{R}^d)$ to $[0,1]^d$.

Our target function f^* is defined by

(2.1)
$$f^{\star} \in \operatorname*{argmin}_{f \in \mathcal{H}^{\alpha}} \widetilde{\mathcal{R}}_{P}(f, \varepsilon).$$

When only a finite sample $\{(X_i, Y_i), i = 1, ..., n\}$ is available, we estimate f^* by minimizing the empirical adversarial risk over a space of estimation functions \mathcal{F}_n , which can vary with n. Specifically, we aim to find an estimator $\hat{f}_n \in \mathcal{F}_n$ that solves

$$\widehat{f}_n \in \operatorname*{argmin}_{f \in \mathcal{F}_n} \widetilde{\mathcal{R}}_{P_n}(f, \varepsilon), \quad \text{where } \widetilde{\mathcal{R}}_{P_n}(f, \varepsilon) = \frac{1}{n} \sum_{i=1}^n \left[\sup_{X'_i \in B_\varepsilon(X_i)} \ell(f(X'_i), Y_i) \right].$$

Here P_n denotes the empirical distribution induced by the samples. The function \hat{f}_n is called an adversarial estimator. Based on the relationship between Lipschitz constraints and adversarial robustness, we focus on the feedforward neural network with constraints on Lipschitz property.

2.5. Feedforward neural networks with norm constraints. A feedforward neural network (FNN) can be represented in the form of

$$g = g_L \circ g_{L-1} \circ \cdots \circ g_0,$$

where $g_i(\boldsymbol{x}) = \sigma(A_i\boldsymbol{x} + \boldsymbol{b}_i)$ and $g_L(\boldsymbol{x}) = A_L\boldsymbol{x}$, with $A_i \in \mathbb{R}^{d_{i+1} \times d_i}$ and $\boldsymbol{b}_i \in \mathbb{R}^{d_{i+1} \times 1}$ for $i = 0, \ldots, L - 1, A_L \in \mathbb{R}^{d_{L+1} \times d_L}$, and $\sigma(\boldsymbol{x}) = \max\{x, 0\}$ being the ReLU activation function (applied componentwise). For simplicity of notation, we use g_{θ} to emphasis that the FNN is parameterized by $\theta = (A_0, \ldots, A_L, \boldsymbol{b}_0, \ldots, \boldsymbol{b}_{L-1})$. The numbers $W = \max\{d_1, \ldots, d_L\}$ and L are called the width and depth of the FNN, respectively. We let $\mathcal{NN}(W, L)$ denote the class of FNNs with width W and depth L. Additionally, we define $\mathcal{NN}(W, L, K)$ as the subset of functions in $\mathcal{NN}(W, L)$ which satisfies the following norm constraint on the weight:

$$\kappa(\theta) := ||A_L|| \prod_{i=0}^{L-1} \max\{||(A_i, \boldsymbol{b}_i)||, 1\} \le K,$$

where the norm satisfies $||A|| = \sup_{||\boldsymbol{x}||_{\infty} \leq 1} ||A\boldsymbol{x}||_{\infty}$. Then for any $g_{\theta} \in \mathcal{NN}(W, L, K)$, we have

$$\operatorname{Lip}(g_{\theta}) \leq \kappa(\theta) \leq K.$$

Therefore, the Lipschitz constants of the functions in $\mathcal{NN}(W, L, K)$ have an uniform upper bound.

3. Nonasymptotic error bounds. In this section, we present our main results of nonasymptotic bounds for the adversarial excess risk. We also discuss the relationship between accuracy and adversarial robustness in the sense that more robustness can lead to less accurate upper bounds for the excess risk.

3.1. Nonasymptotic error bounds for adversarial excess risk. The adversarial estimator based on FNNs with norm constraints is defined by

(3.1)
$$\widehat{f}_n \in \operatorname*{argmin}_{f \in \mathcal{NN}(W,L,K)} \widetilde{\mathcal{R}}_{P_n}(f,\varepsilon).$$

We evaluate its performance via the adversarial excess risk

$$\widetilde{\mathcal{E}}(\widehat{f}_n,\varepsilon) = \widetilde{\mathcal{R}}_P(\widehat{f}_n,\varepsilon) - \inf_{f \in \mathcal{H}^\alpha \cup \mathcal{NN}(W,L,K)} \widetilde{\mathcal{R}}_P(f,\varepsilon)$$

The adversarial excess risk is nonnegative. It is a measure for evaluating the performance of an adversarial estimator for future data using the optimal population adversarial risk as a benchmark.

To investigate the adversarial excess risk, we show it can be decomposed into (3.2), where \mathcal{E}_{gen} represents the generalization error, which is the difference between the population adversarial risk and empirical adversarial risk, and \mathcal{E}_{app} represents the approximation error due to model misspecification, which measures the distance between the target function and the space of estimation functions that may not contain the target function. By investigating both \mathcal{E}_{qen} and \mathcal{E}_{app} , we establish a nonasymptotic error bound on the adversarial excess risk.

Theorem 3.1. Consider a Hölder space \mathcal{H}^{α} with $\alpha = r + \beta \geq 1$, where $r \in \mathbb{N}_0$ and $\beta \in (0,1]$. Let $\gamma = \lceil \log_2(d+r) \rceil$. Suppose Assumptions 2.1–2.2 hold. Suppose $W \geq c(K/\log^{\gamma} K)^{(2d+\alpha)/(2d+2)}$ for a constant c > 0, and $L \geq 4\gamma + 2$. Then for any adversarial estimator \hat{f}_n in (3.1) and adversarial attack level $\varepsilon > 0$, we have

(3.2)
$$\mathcal{E}(f_n,\varepsilon) \lesssim \mathcal{E}_{gen} + \mathcal{E}_{app},$$

where

$$\mathcal{E}_{gen} = K\varepsilon n^{-1} + WL\sqrt{\log(W^2L)}n^{-1/2}\sqrt{\log n} + n^{-\min\{1/2,\alpha/d\}}\log^{c(\alpha,d)}n,$$

$$\mathcal{E}_{app} = (K/\log^{\gamma}K)^{-\alpha/(d+1)}.$$

Here $c(\alpha, d) = 1$ when $d = 2\alpha$, and $c(\alpha, d) = 0$ otherwise. If we further select $K \simeq n^{(d+1)/(2d+3\alpha)}$ and $WL \simeq n^{(2d+\alpha)/(4d+6\alpha)}$, then we have

(3.3)
$$\widetilde{\mathcal{E}}(\widehat{f}_n,\varepsilon) \lesssim n^{-(d+3\alpha-1)/(2d+3\alpha)}\varepsilon + n^{-\alpha/(2d+3\alpha)}\log n^{\xi},$$

where $\xi = \max\{1, \gamma \alpha / (d+1)\}.$

The upper bound (3.2) is determined by the smoothness property of the Hölder space, the structure of the estimation function class $\mathcal{NN}(W, L, K)$, the adversarial attack level ε , and sample size n. There is a trade-off between the two errors. Specifically, \mathcal{E}_{gen} increases with the complexity of $\mathcal{NN}(W, L, K)$, with larger W, L, and K leading to a larger upper bound. On the other hand, as K becomes larger, the error \mathcal{E}_{app} decreases. To achieve the best error rate, we balance the trade-off between the two errors and show the rate can reach $n^{-\alpha/(2d+3\alpha)}$ up to a logarithmic factor for suitable chosen ε .

The proposed analysis method is applicable to the settings with different models, loss functions, estimation function classes, and adversarial attacks. For example, we can apply the method to the setting where a general ℓ_r adversarial attack is used. The upper bounds on the corresponding generalization error and approximation error can be obtained based on the results on Rademacher complexity of general adversarial loss function classes [3, 31] and the results on approximation power of different estimation function classes such as deep neural networks [51, 26, 22]. In Theorem 3.1, we employ the approximation error theory as established in [22], specifically developed for norm-constrained FNNs.

The result (3.3) shows how the bounds for the adversarial excess risk depends on the adversarial attack level ε , where ε is allowed to vary with n. Let $e_n = n^{(d+2\alpha-1)/(2d+3\alpha)}$. When $\varepsilon = O(e_n)$, the error rate can reach $n^{-\alpha/(2d+3\alpha)}$ up to a logarithmic factor. However, if ε grows faster than e_n , the error rate of $\widetilde{\mathcal{E}}(\widehat{f}_n, \varepsilon)$ is dominated by the rate of ε . Moreover, the convergence of $\widetilde{\mathcal{E}}(\widehat{f}_n, \varepsilon)$ cannot be guaranteed when ε grows faster than $n^{(d+3\alpha-1)/(2d+3\alpha)}$.

As mentioned above, the error rate of the adversarial excess risk can reach $n^{-\alpha/(2d+3\alpha)}$ up to a logarithmic factor when ε is appropriately selected. Here we only require the Lipschitz property of the loss function. We will further show in section 4.2 that the error rate can be improved to $n^{-2\alpha/(2d+5\alpha)}$ up to a logarithmic factor when using the quadratic loss, where the improvement is due to an improved approximation error bound.

Some recent papers have studied the convergence properties of deep neural network under the excess risk (3.4), where the data are naturally unperturbed [39, 7, 21]. The results are Table 3

Comparison of nonasymptotic error bounds between our result and some related results (up to a logarithmic factor).

	Setup	Estimation function class	Error bound
[39, 7, 21]	natural	FNNs	$n^{-2\alpha/(d+2\alpha)}$
[22]	natural	FNNs with norm constraints	$n^{-\alpha/(d+2\alpha+1)}$
This paper	adversarial	FNNs with norm constraints	$n^{-2\alpha/(2d+5\alpha)}$

generally established under a certain smoothness assumption on the target function. And it is typically assumed that the target function is in a Hölder class with a smoothness index α . The results show that the deep neural network estimation could achieve the optimal minimax rate $n^{-2\alpha/(d+2\alpha)}$ established by [41]. Though the structures of neural networks vary in these works, which include different choices of width, depth, and activation functions, they make no constraint on the Lipschitz property of neural networks. [22] investigated the approximation properties of FNNs with norm constraints. Intuitively, a norm-constrained neural network class would be smaller in size compared to an unconstrained neural network class with the same structure. Therefore, the benefit of the Lipschitz property comes at the cost of losing the approximation power, which would lead to larger approximation errors. This is demonstrated in [22], where the error rate of the excess risk only reaches the rate $n^{-\alpha/(d+2\alpha+1)}$ up to a logarithmic factor. The above discussion is summarized in Table 3.

To date, existing precise lower bound results for the adversarial settings have been established based on strong assumptions, where the parametric model and Gaussian distribution are commonly assumed. Specifically, [19, 20] derived expressions of the adversarial risks in the linear regression and parametric binary classification problems under the Gaussian assumption. However, for the general nonparametric adversarial setting, there is no established lower minimax bound result, and the problem becomes much harder as precise expression of adversarial risk is not available. Since the natural risk is upper bounded by the adversarial risk, the minimax optimal rate for the natural risk naturally becomes applicable to establishing a lower bound of adversarial risk. Therefore, $n^{-2\alpha/(d+2\alpha)}$ is a valid but trivial lower bound for nonparametric adversarial settings. However, the lower minimax bound of adversarial risk is likely to depend on the adversarial attack level ε . Moreover, the class of norm-constrained FNNs is smaller than the class of FNNs used to establish the rate $n^{-2\alpha/(d+2\alpha)}$ of natural risk. Therefore, it seems reasonable to conjecture that the lower minimax bound for the adversarial risk cannot reach $n^{-2\alpha/(d+2\alpha)}$ and will vary with ε . In [22], a lower bound for the approximation error of norm-constrained FNNs was established. However, after much work, it appears that this result does not lead to a lower minimax bound of adversarial risk. Consequently, the tightness of our established upper bound and a rigorous derivation of the lower minimax bound still need to be explored. This is an interesting and challenging problem that deserves further study in the future.

3.2. Evaluation of adversarial estimator under some other risks. We also evaluate the performance of the adversarial estimator under some other risks. We first consider the natural risk and study the excess risk defined by

(3.4)
$$\mathcal{E}(\widehat{f}_n) = \mathcal{R}_P(\widehat{f}_n) - \inf_{f \in \mathcal{H}^{\alpha}} \mathcal{R}_P(f).$$

Corollary 3.2. Suppose the conditions of Theorem 3.1 are satisfied and $\alpha \geq 1$. Then for any adversarial estimator \hat{f}_n in (3.1), we have

$$\mathcal{E}(\widehat{f}_n) \lesssim n^{-\alpha/(2d+3\alpha)} \log n^{\xi} + \varepsilon,$$

where $\xi = \max\{1, \gamma \alpha/(d+1)\}.$

Corollary 3.2 shows that the upper bound for the excess risk of the adversarial estimator is not guaranteed to converge. The increase in the upper bound for the excess risk becomes significant when the adversarial robustness reaches a certain level. Previous studies have mostly focused on specific scenarios and made certain assumptions on the data distribution. For instance, [19, 20] analyzed the trade-offs in linear regression and binary classification with linear classifier, assuming the data was normally distributed. However, a comprehensive analysis of this problem is still lacking. We provide an upper bound for the excess risk that increases with the adversarial robustness level. Our result sheds light on the theoretical understanding of the trade-offs, but a complete analysis requires lower bounds for the excess risk.

We now consider the local worst-case risk. Specifically, the local worst-case risk with 1st Wasserstein distance is defined by

$$\mathcal{R}_{P,1}(f,\varepsilon) = \sup_{Q:W_1(Q,P) \le \varepsilon} \mathcal{R}_Q(f),$$

where the distribution Q runs over an uncertainty set around the data generating distribution P. The excess risk with respect to the local worst-case risk is defined by

$$\mathcal{E}_1(\widehat{f}_n,\varepsilon) = \mathcal{R}_{P,1}(\widehat{f}_n,\varepsilon) - \inf_{f \in \mathcal{H}^\alpha \cup \mathcal{NN}(W,L,K)} \mathcal{R}_{P,1}(f,\varepsilon).$$

Corollary 3.3. Suppose the conditions of Theorem 3.1 are satisfied and $\operatorname{Lip}(\ell) < \infty$. Then for any adversarial estimator \hat{f}_n in (3.1), we have

$$\mathcal{E}_1(\widehat{f}_n,\varepsilon) \lesssim n^{-\alpha/(2d+3\alpha)} \log n^{\xi} + K\varepsilon,$$

where $\xi = \max\{1, \gamma \alpha / (d+1)\}.$

4. Examples. In this section, we consider the more specific settings of classification and regression and apply Theorem 3.1 to classification and regression problems.

4.1. Classification. Suppose that (X, Y) follows an unknown distribution P on $\mathcal{X} \times \{-1, 1\}$. A basic goal of binary classification is to predict the label Y, when we only observe a predictor X in a random pair $(X, Y) \sim P$. A commonly used loss function is the classification loss $\ell_{\text{class}} : \mathbb{R} \times \{-1, 1\} \mapsto [0, \infty)$, defined by $\ell_{\text{class}}(u, y) = \mathbf{1}\{\text{sign}(u)y \leq 0\}$, where sign(u) = 1 when $u \geq 0$, and sign(u) = -1 otherwise. Let $f : \mathcal{X} \mapsto \mathbb{R}$ be a score function, and let the associated binary classifier be $\text{sign}f(\cdot)$. The natural classification risk and the adversarial classification risk of the score function f are

856

$$\begin{aligned} \mathcal{R}_{\mathrm{class},P}(f) &= \mathbb{E}_{(X,Y)\sim P} \mathbf{1}\{\mathrm{sign} f(X) \neq Y\}, \\ \widetilde{\mathcal{R}}_{\mathrm{class},P}(f,\varepsilon) &= \mathbb{E}_{(X,Y)\sim P} \left[\sup_{X' \in B_{\varepsilon}(X)} \mathbf{1}\left\{\mathrm{sign} f(X') \neq Y\right\} \right] \end{aligned}$$

Let $\eta(\boldsymbol{x}) = P(Y = 1 | X = \boldsymbol{x})$. Define $c_{\varepsilon}(\boldsymbol{x}, \boldsymbol{x}') = \mathbf{1}\{\|\boldsymbol{x} - \boldsymbol{x}'\|_{\infty} > 2\varepsilon\}$ and let the corresponding optimal transport cost D_{ε} be defined by

$$D_{\varepsilon}(P,Q) = \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_{(X_1,X_2) \sim \pi}[c_{\varepsilon}(X_1,X_2)].$$

The minimum value of the natural classification risk is given by

$$\mathcal{R}_{\text{class},P}^{\star} = \inf_{f \text{ measurable}} \mathcal{R}_{\text{class},P}(f) = \mathbb{E}[\min\{\eta(X), 1 - \eta(X)\}],$$

which is reached when f is the Bayes classifier, i.e., $f(\mathbf{x}) = \text{sign}(2\eta(\mathbf{x}) - 1)$ [40]. The minimum value of the adversarial classification risk can be expressed as

$$\widetilde{\mathcal{R}}_{\text{class},P}^{\star}(\varepsilon) = \inf_{f \text{ measurable}} \widetilde{\mathcal{R}}_{\text{class},P}(f,\varepsilon) = \frac{1}{T+1} \left[1 - \inf_{Q \in \mathcal{P}(\mathcal{X}): Q \preceq TP_0} D_{\varepsilon}(Q,P_1) \right],$$

where $P_1 = P_{X|Y=1}$, $P_0 = P_{X|Y=-1}$, and T = P(Y = -1)/P(Y = 1) [35, Theorem 6.2].

The natural classification loss and its adversarial counterpart are nonsmooth and nonconvex. Many surrogate losses have been considered in the context of standard classification. We specifically focus on margin-based loss, where a margin loss function ϕ exists such that the loss function satisfies $\ell(u, y) = \phi(uy)$, $(u, y) \in \mathbb{R} \times \{-1, 1\}$. In general, the margin loss is selected to have a property called consistency, which is satisfied by a large family of convex losses [40]. However, the adversarial version of these margin losses may not show the same consistency properties with respect to the adversarial classification loss. Moreover, [29] showed that no convex margin loss can be calibrated in the adversarial setting. Consequently, it is challenging to study consistency in the general adversarial setting.

Let $C_{\text{class}}(\eta, \boldsymbol{x}, f) = \mathbf{1}\{f(\boldsymbol{x}) < 0\}\eta + \mathbf{1}\{f(\boldsymbol{x}) \ge 0\}(1-\eta) \text{ and } C^{\star}_{\text{class}}(\eta, \boldsymbol{x}) = \min\{\eta, 1-\eta\}.$ Let $C_{\phi}(\eta, \boldsymbol{x}, f) = \phi(f(\boldsymbol{x}))\eta + \phi(-f(\boldsymbol{x}))(1-\eta) \text{ and } C^{\star}_{\phi}(\eta, \boldsymbol{x}) = \inf_{\alpha} \phi(\alpha)\eta + \phi(-\alpha)(1-\eta).$ Define $\mathcal{R}_{P}^{\star} = \inf_{f \text{ measurable }} \mathcal{R}_{P}(f).$

Assumption 4.1. For any $\eta \in [0,1]$, $\boldsymbol{x} \in \mathcal{X}$ and measurable function f,

$$C_{\phi}(\eta, \boldsymbol{x}, f) - C_{\phi}^{\star}(\eta, \boldsymbol{x}) \ge a(C_{\text{class}}(\eta, \boldsymbol{x}, f) - C_{\text{class}}^{\star}(\eta, \boldsymbol{x}))$$

holds for a positive constant a.

Assumption 4.2. There exist positive constants c and b such that

$$\phi(0) - C^{\star}_{\phi}(\eta, \boldsymbol{x}) \ge b(1 - C^{\star}_{\text{class}}(\eta, \boldsymbol{x}))$$

when $|\eta - 1/2| > c$.

Assumptions 4.1 and 4.2 can be satisfied by some common margin losses, such as the hinge loss.

Corollary 4.3. Suppose the conditions of Theorem 3.1 are satisfied and ϕ is a continuous decreasing margin function satisfying Assumptions 4.1 and 4.2. Assume $|\eta(\boldsymbol{x}) - 1/2| > c$ for any $\boldsymbol{x} \in \mathcal{X}$ and $\inf_{f \in \mathcal{H}^{\alpha}} \mathcal{R}_{P}(f) = \mathcal{R}_{P}^{\star}$. Then

$$\widetilde{\mathcal{R}}_{\mathrm{class},P}(\widehat{f}_n,\varepsilon) - \widetilde{\mathcal{R}}^{\star}_{\mathrm{class},P}(\varepsilon) \lesssim n^{-\alpha/(2d+3\alpha)} \log n^{\xi} + \varepsilon.$$

For the case where there might be $\eta(\boldsymbol{x}) = 1/2$, we show that the natural classification risk of the adversarial estimator converges to $\mathcal{R}^{\star}_{\text{class},P}$ when ε goes to 0.

Corollary 4.4. Suppose the conditions of Theorem 3.1 are satisfied and ϕ is a margin function satisfying Assumptions 4.1. Assume $\inf_{f \in \mathcal{H}^{\alpha}} \mathcal{R}_{P}(f) = \mathcal{R}_{P}^{\star}$. Then

$$\mathcal{R}_{\text{class},P}(\widehat{f}_n) - \mathcal{R}^{\star}_{\text{class},P} \lesssim n^{-\alpha/(2d+3\alpha)} \log n^{\xi} + \varepsilon.$$

4.2. Regression. Consider a nonparametric regression model

$$(4.1) Y = f_0(X) + \eta,$$

where Y is a response, $X \in \mathcal{X} \subseteq [0,1]^d$ is a d-dimensional covariate vector, $f_0 \in \mathcal{H}^{\alpha}$ is an unknown regression function, and η is an unobservable error satisfying $\mathbb{E}(\eta|X) = 0$ and $\mathbb{E}(\eta^2) < \infty$. Under model (4.1) and the quadratic loss $\ell(u, y) = (u - y)^2$, we denote the corresponding adversarial estimator (3.1) by \hat{f}_n^{ls} . We measure the distance between \hat{f}_n^{ls} and f_0 using the $L_2(P)$ -norm $\|\cdot\|_2 := \|\cdot\|_{L_2(P_X)}$, that is, $\|f\|_2 = \sqrt{\mathbb{E}|f(X)|^2}$. To relax the boundedness requirement for Y, we assume that Y follows a subexponential distribution.

Assumption 4.5. There exists a constant $\sigma_Y > 0$ such that $\mathbb{E}[\exp\{\sigma_Y|Y|\}] < \infty$.

First, we establish a new error bound for the adversarial excess risk when utilizing the quadratic loss.

Theorem 4.6. Consider a Hölder class \mathcal{H}^{α} with $\alpha = r + \beta \geq 1$, where $r \in \mathbb{N}_0$ and $\beta \in (0,1]$. Let $\gamma = \lceil \log_2(d+r) \rceil$. Suppose Assumptions 2.1 and 4.5 hold. Suppose $W \geq c(K/\log^{\gamma} K)^{(2d+\alpha)/(2d+2)}$ for a constant c > 0 and $L \geq 4\gamma + 2$. If we select $K \approx n^{(d+1)/(2d+5\alpha)}$ and $WL \approx n^{(2d+\alpha)/(4d+10\alpha)}$, then for any adversarial estimator \widehat{f}_n^{ls} satisfying $\|\widehat{f}_n^{ls}\|_{\infty} \leq M_n$ with M_n growing at rate $\log n$, we have

$$\mathbb{E}[\widetilde{\mathcal{E}}(\widehat{f}_n^{ls},\varepsilon)] \lesssim n^{-2\alpha/(2d+5\alpha)} \log n^{\lambda} + n^{(d+1)/(2d+5\alpha)} \log n\varepsilon,$$

where $\lambda = \max\{4, 2\gamma\alpha/(d+1)\}.$

Theorem 4.6 shows that the error rate of the adversarial excess risk of \hat{f}_n^{ls} can reach $n^{-2\alpha/(2d+5\alpha)}$ up to a logarithmic factor when $\varepsilon = O(n^{-(d+2\alpha+1)/(2d+5\alpha)})$. This improves the rate $n^{-\alpha/(2d+3\alpha)}$ given by Theorem 3.1. This is because a better control of the approximation error can be obtained with the quadratic loss function. We also obtain the convergence rate of $\|\hat{f}_n^{ls} - f_0\|_2$.

Corollary 4.7. Suppose the conditions of Theorem 4.6 are satisfied. Then

$$\mathbb{E} \| \hat{f}_n^{ls} - f_0 \|_2^2 \lesssim n^{-2\alpha/(2d+5\alpha)} \log n^{\lambda} + n^{(d+1)/(2d+5\alpha)} \log n\varepsilon.$$

If further $\varepsilon = O(n^{-(d+2\alpha+1)/(2d+5\alpha)})$, then

$$\mathbb{E}\|\widehat{f}_n^{ls} - f_0\|_2^2 \lesssim n^{-2\alpha/(2d+5\alpha)}\log n^{\lambda}$$

5. Related work. There is a line of work focusing on the analysis of the Rademacher complexity of adversarial loss function class [52, 23, 3, 31]. Specifically, [52] investigated the adversarial Rademacher complexity of the linear models under perturbations measured in the ℓ_{∞} -norm. The result was later generalized by [23] and [3] to the cases where the perturbations were measured in a general ℓ_r -norm. For neural network models, [52, 3] investigated adversarial training when the model was a neural network with a single hidden layer. In [23] and [31], deep neural networks were studied. [23] proposed a tree transformation to upper bound the adversarial loss function. In [31], the covering number of the adversarial loss function class over an extended training set. The Rademacher complexity of the adversarial loss function class. [25] and [46] transformed the adversarial learning problem into a DRO problem. However, in all the aforementioned works, the authors did not consider the approximation error and only focused on the generalization error; see Table 1.

The problem of the trade-offs between accuracy and adversarial robustness has been studied recently [28, 38, 44, 37, 15, 27, 49, 14]. The works [19] and [20] gave a precise theoretical characterization of the trade-offs in the linear regression and parametric binary classification problems under the Gaussian assumption. For the adversarial training, [20] characterized its trade-off curve by calculating the natural risk and adversarial risk of the adversarial estimator that was derived from different adversarial attack levels. They found that the adversarial training hurt the accuracy if robustness is pursued. However, there is still a lack of systematic theoretical understanding of the trade-offs in general nonparametric settings.

The vulnerability of typical neural networks to attack often stems from their lack of Lipschitzness, where small adversarial perturbations in the input can result in significant perturbations in the output [43]. While the Lipschitz condition has been extensively studied in adversarial training, several relaxations of this assumption have been explored in the literature. These include local Lipschitzness [50] and the local cross-Lipschitz condition [18, 48]. However, it's important to note that these works did not consider theoretical analyses regarding the generalization property.

6. Conclusions. In this paper, we have proposed a general approach to evaluating the generalization performance of the estimators based on adversarial training under misspecified models. The adversarial risk is shown to be equivalent to the risk induced by a distributional adversarial attack under certain smoothness conditions. This shows that the adversarial training procedure is well-defined. We have established nonasymptotic error bounds on the adversarial excess risk, which achieve the rate $O(n^{-\alpha/(2d+3\alpha)})$ up to a logarithmic factor for a Lipschitz loss function and can be improved to $O(n^{-2\alpha/(2d+5\alpha)})$ up to a logarithmic factor when using the quadratic loss.

There are several interesting problems that deserve further study. First, the Lipschitztype condition (e.g., Assumption 2.2) plays an important role in our analysis. However, this assumption is not satisfied in many practical cases. It would be interesting to relax this assumption in the future. Also, we have only considered robustness against adversarial examples in X. How to generalize the results to the case when there are also adversarial examples in both X and Y is an important problem for future work. Finally, a complete analysis of the trade-offs between accuracy and adversarial robustness requires the establishment of lower bounds for the adversarial excess risk.

Appendix. In this appendix, we give the proof of our main result, Theorem 3.1. Proofs of the other results and additional technical details are given in the supplementary material (supplement.pdf [local/web 358KB]).

Appendix A. Proof of Theorem 3.1. Let $f_0 \in \operatorname{argmin}_{f \in \mathcal{NN}(W,L,K)} \widetilde{\mathcal{R}}_P(f,\varepsilon)$. Then we have

$$\widetilde{\mathcal{E}}(\widehat{f}_n,\varepsilon) = \widetilde{\mathcal{R}}_P(\widehat{f}_n,\varepsilon) - \inf_{f \in \mathcal{H}^\alpha \cup \mathcal{NN}(W,L,K)} \widetilde{\mathcal{R}}_P(f,\varepsilon) = \widetilde{\mathcal{R}}_P(\widehat{f}_n,\varepsilon) - \min\left\{\widetilde{\mathcal{R}}_P(f^\star,\varepsilon), \widetilde{\mathcal{R}}_P(f_0,\varepsilon)\right\} = \max\{\mathcal{E}_1,\mathcal{E}_2\},$$

where $\mathcal{E}_1 = \widetilde{\mathcal{R}}_P(\widehat{f}_n, \varepsilon) - \widetilde{\mathcal{R}}_P(f^*, \varepsilon)$ and $\mathcal{E}_2 = \widetilde{\mathcal{R}}_P(\widehat{f}_n, \varepsilon) - \widetilde{\mathcal{R}}_P(f_0, \varepsilon)$. For the error \mathcal{E}_1 , we have the decomposition

$$\mathcal{E}_{1} = \mathcal{R}_{P}(f_{n},\varepsilon) - \mathcal{R}_{P_{n}}(f_{n},\varepsilon) + \mathcal{R}_{P_{n}}(f_{n},\varepsilon) - \mathcal{R}_{P_{n}}(f^{\star},\varepsilon) + \widetilde{\mathcal{R}}_{P_{n}}(f^{\star},\varepsilon) - \widetilde{\mathcal{R}}_{P}(f^{\star},\varepsilon) = I_{1} + I_{2} + I_{3},$$

where the errors are $I_1 = \widetilde{\mathcal{R}}_P(\widehat{f}_n, \varepsilon) - \widetilde{\mathcal{R}}_{P_n}(\widehat{f}_n, \varepsilon), I_2 = \widetilde{\mathcal{R}}_{P_n}(\widehat{f}_n, \varepsilon) - \widetilde{\mathcal{R}}_{P_n}(f^*, \varepsilon)$, and $I_3 = \widetilde{\mathcal{R}}_{P_n}(f^*, \varepsilon) - \widetilde{\mathcal{R}}_P(f^*, \varepsilon)$. For the error \mathcal{E}_2 , based on $\widetilde{\mathcal{R}}_{P_n}(\widehat{f}_n, \varepsilon) \leq \widetilde{\mathcal{R}}_{P_n}(f_0, \varepsilon)$, we have the decomposition

$$\begin{aligned} \mathcal{E}_{2} &= \widetilde{\mathcal{R}}_{P}(\widehat{f}_{n},\varepsilon) - \widetilde{\mathcal{R}}_{P_{n}}(\widehat{f}_{n},\varepsilon) + \widetilde{\mathcal{R}}_{P_{n}}(\widehat{f}_{n},\varepsilon) - \widetilde{\mathcal{R}}_{P_{n}}(f_{0},\varepsilon) \\ &+ \widetilde{\mathcal{R}}_{P_{n}}(f_{0},\varepsilon) - \widetilde{\mathcal{R}}_{P}(f_{0},\varepsilon) \\ &\leq I_{1} + I_{4}, \end{aligned}$$

where $I_4 = \widetilde{\mathcal{R}}_{P_n}(f_0, \varepsilon) - \widetilde{\mathcal{R}}_P(f_0, \varepsilon)$. We show the nonasymptotic upper bound for the adversarial excess risk by deriving N upper bounds for each error term.

A.1. Bounds for I_1 and I_4. For any $f \in \mathcal{NN}(W, L, K)$ and $z = (x, y) \in \mathcal{Z}$, define

$$\tilde{\ell}(f, \boldsymbol{z}) = \sup_{\boldsymbol{x}' \in B_{\varepsilon}(\boldsymbol{x})} \ell(f(\boldsymbol{x}'), y) = \sup_{\boldsymbol{\delta} \in B_{\varepsilon}(0)} \ell(f(\boldsymbol{x} + \boldsymbol{\delta}), y).$$

Since \widehat{f}_n and f_0 belong to the class $\mathcal{NN}(W, L, K)$, we have

$$I_1 = \widetilde{\mathcal{R}}_P(\widehat{f}_n, \varepsilon) - \widetilde{\mathcal{R}}_{P_n}(\widehat{f}_n, \varepsilon) \le \sup_{f \in \mathcal{NN}(W, L, K)} \{ \mathbb{E}_{Z \sim P} \widetilde{\ell}(f, Z) - \mathbb{E}_{Z \sim P_n} \widetilde{\ell}(f, Z) \}$$

and

$$I_4 = \widetilde{\mathcal{R}}_{P_n}(f_0, \varepsilon) - \widetilde{\mathcal{R}}_P(f_0, \varepsilon) \le \sup_{f \in \mathcal{NN}(W, L, K)} \{ \mathbb{E}_{Z \sim P_n} \widetilde{\ell}(f, Z) - \mathbb{E}_{Z \sim P} \widetilde{\ell}(f, Z) \}.$$

Let the random vector $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_n)$ consist of i.i.d. Rademacher variables that are independent of the data. The Rademacher variable takes equal probability of being 1 or -1.

Denote the samples by $Z_{1:n} = \{Z_i\}_{i=1}^n$, with $Z_i = (X_i, Y_i)$. Let $Z'_i = (X'_i, Y'_i), i = 1, ..., n$, be generated i.i.d. from P and be independent of $Z_{1:n}$. The sample $Z'_{1:n} = \{Z'_i\}_{i=1}^n$ is called as the ghost sample of $Z_{1:n}$. Then

(A.1)

$$\begin{split}
\sup_{f \in \mathcal{NN}(W,L,K)} \left\{ \mathbb{E}_{P}[\tilde{\ell}(f,Z)] - \mathbb{E}_{P_{n}}[\tilde{\ell}(f,Z)] \right\} \\
&= \sup_{f \in \mathcal{NN}(W,L,K)} \mathbb{E}_{\sigma} \left\{ \mathbb{E}_{P}[\tilde{\ell}(f,Z)] - \mathbb{E}_{P_{n}}[\tilde{\ell}(f,Z)] \right\} \\
&= \sup_{f \in \mathcal{NN}(W,L,K)} \mathbb{E}_{\sigma} \left\{ \mathbb{E}_{Z'_{1:n}} \left[\frac{1}{n} \sum_{i=1}^{n} \tilde{\ell}(f,Z'_{i}) \right] - \frac{1}{n} \sum_{i=1}^{n} \tilde{\ell}(f,Z_{i}) \right] \right\} \\
&= \sup_{f \in \mathcal{NN}(W,L,K)} \mathbb{E}_{Z'_{1:n}} \left\{ \mathbb{E}_{\sigma} \left[\frac{1}{n} \sum_{i=1}^{n} \tilde{\ell}(f,Z'_{i}) - \tilde{\ell}(f,Z_{i}) \right] \right\} \\
&\leq \mathbb{E}_{Z'_{1:n}} \mathbb{E}_{\sigma} \left\{ \sup_{f \in \mathcal{NN}(W,L,K)} \left[\frac{1}{n} \sum_{i=1}^{n} \sigma_{i} \left[\tilde{\ell}(f,Z'_{i}) - \tilde{\ell}(f,Z_{i}) \right] \right\} ,
\end{split}$$

where the last equality holds since $\tilde{\ell}(f, Z'_i) - \tilde{\ell}(f, Z_i)$ are symmetric random variables, for which they have the same distribution as $\sigma_i(\tilde{\ell}(f, Z'_i) - \tilde{\ell}(f, Z_i))$ [47, Chapter 6.4]. Define the class of functions \mathcal{L}_n by

$$\mathcal{L}_n = \Big\{ \tilde{\ell}(f, \boldsymbol{z}) : \boldsymbol{\mathcal{Z}} \mapsto \mathbb{R} \mid f \in \mathcal{NN}(W, L, K) \Big\}.$$

For a given set of samples z_1, \ldots, z_n from \mathcal{Z} , the empirical Rademacher complexity of class \mathcal{L}_n is defined by

$$\widehat{\Re}_n(\mathcal{L}_n) = \mathbb{E}_{\boldsymbol{\sigma}} \left\{ \sup_{f \in \mathcal{NN}(W,L,K)} \frac{1}{n} \sum_{i=1}^n \sigma_i \widetilde{\ell}(f, \boldsymbol{z}_i) \right\}.$$

We analyze the Rademacher complexity following the method motivated by [31]. For a given $\tau \in (0, \varepsilon)$, let $C_{B_{\varepsilon}}(\tau)$ be a $(\tau, \|\cdot\|_{\infty})$ -cover of $B_{\varepsilon}(0)$ with the smallest cardinality M_{τ} . Denote the elements of $C_{B_{\varepsilon}}(\tau)$ by $\delta_1, \ldots, \delta_{M_{\tau}}$. It follows by Lemma SM1.3 that $\log M_{\tau} \leq cd \log(\varepsilon \tau^{-1})$ for a constant c. For any $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{Z}$, the continuity of ℓ and f implies that there exists $\boldsymbol{\delta}' \in B_{\varepsilon}(0)$ such that $\tilde{\ell}(f, \boldsymbol{z}) = \ell(f(\boldsymbol{x} + \boldsymbol{\delta}'), \boldsymbol{y})$. Then

$$\begin{split} \tilde{\ell}(f, \boldsymbol{z}) &- \max_{j} \ell(f(\boldsymbol{x} + \boldsymbol{\delta}_{j}), y) = \ell(f(\boldsymbol{x} + \boldsymbol{\delta}'), y) - \max_{j} \ell(f(\boldsymbol{x} + \boldsymbol{\delta}_{j}), y) \\ &\leq \min_{j} |\ell(f(\boldsymbol{x} + \boldsymbol{\delta}'), y) - \ell(f(\boldsymbol{x} + \boldsymbol{\delta}_{j}), y)| \\ &\leq \operatorname{Lip}^{1}(\ell) \operatorname{Lip}(f) \min_{j} \|\boldsymbol{\delta}' - \boldsymbol{\delta}_{j}\|_{\infty} \\ &\leq \operatorname{Lip}^{1}(\ell) \operatorname{Lip}(f) \tau. \end{split}$$

Therefore, for any $f \in \mathcal{NN}(W, L, K)$,

$$\begin{split} &\frac{1}{n}\sum_{i=1}^{n}\sigma_{i}\tilde{\ell}(f,\boldsymbol{z}_{i}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\sigma_{i}\tilde{\ell}(f,\boldsymbol{z}_{i}) - \sigma_{i}\max_{j}\ell(f(\boldsymbol{x}_{i}+\boldsymbol{\delta}_{j}),y_{i}) + \sigma_{i}\max_{j}\ell(f(\boldsymbol{x}_{i}+\boldsymbol{\delta}_{j}),y_{i})\right\} \\ &\leq \mathrm{Lip}^{1}(\ell)K\tau + \frac{1}{n}\sum_{i=1}^{n}\left\{\sigma_{i}\max_{j}\ell(f(\boldsymbol{x}_{i}+\boldsymbol{\delta}_{j}),y_{i})\right\}. \end{split}$$

This leads to an upper bound of $\widehat{\Re}_n(\mathcal{L}_n)$ as follows:

$$\widehat{\Re}_n(\mathcal{L}_n) \leq \mathbb{E}_{\boldsymbol{\sigma}} \left\{ \sup_{f \in \mathcal{NN}(W,L,K)} \frac{1}{n} \sum_{i=1}^n \sigma_i \max_j \ell(f(\boldsymbol{x}_i + \boldsymbol{\delta}_j), y_i) \right\} + \operatorname{Lip}^1(\ell) K \tau.$$

Define the class

$$\mathcal{L}_{n,\tau} = \left\{ \max_{j} \ell(f(\boldsymbol{x} + \boldsymbol{\delta}_{j}), y) : \mathcal{Z} \mapsto \mathbb{R} \mid f \in \mathcal{NN}(W, L, K) \right\}.$$

Let $\mathcal{N}(u, \mathcal{L}_{n,\tau}, L_{\infty}(P_n))$ denote the covering number of the class $\mathcal{L}_{n,\tau}$ under the data-dependent L_{∞} metric. Define $S_{nM_{\tau}} = \{\boldsymbol{x}_i + \boldsymbol{\delta}_j : i = 1, ..., n, j = 1, ..., M_{\tau}\}$. For the dataset $S_{nM_{\tau}}$, let $\mathcal{N}(u, \mathcal{N}\mathcal{N}(W, L, K), L_{\infty}(P_{nM_{\tau}}))$ denote the covering number of the class $\mathcal{N}\mathcal{N}(W, L, K)$ under the data-dependent L_{∞} metric. For any $f \in \mathcal{N}\mathcal{N}(W, L, K)$, there exists f' such that $\max_{i,j} |f(\boldsymbol{x}_i + \boldsymbol{\delta}_j) - f'(\boldsymbol{x}_i + \boldsymbol{\delta}_j)| \leq u$, which leads to

$$\begin{split} \max_{i} &|\max_{j} \ell(f(\boldsymbol{x}_{i} + \boldsymbol{\delta}_{j}), y_{i}) - \max_{j} \ell(f'(\boldsymbol{x}_{i} + \boldsymbol{\delta}_{j}), y_{i})| \\ &\leq \max_{i,j} |\ell(f(\boldsymbol{x}_{i} + \boldsymbol{\delta}_{j}), y_{i}) - \ell(f'(\boldsymbol{x}_{i} + \boldsymbol{\delta}_{j}), y_{i})| \\ &\leq \operatorname{Lip}^{1}(\ell) \max_{i,j} |f(\boldsymbol{x}_{i} + \boldsymbol{\delta}_{j}) - f'(\boldsymbol{x}_{i} + \boldsymbol{\delta}_{j})| \\ &\leq \operatorname{Lip}^{1}(\ell) u. \end{split}$$

Hence, we show

$$\mathcal{N}(u, \mathcal{L}_{n,\tau}, L_{\infty}(P_n)) \leq \mathcal{N}(u/\operatorname{Lip}^1(\ell), \mathcal{N}\mathcal{N}(W, L, K), L_{\infty}(P_{nM_{\tau}})).$$

Suppose the functions in $\mathcal{NN}(W, L, K)$ are uniformly bounded; otherwise they can be truncated. Define the uniform covering number of $\mathcal{NN}(W, L, K)$ by

$$\mathcal{N}_{\infty}(u, \mathcal{N}\mathcal{N}(W, L, K), n) = \sup_{P_n} \mathcal{N}(u, \mathcal{N}\mathcal{N}(W, L, K), L_{\infty}(P_n)),$$

where the supremum runs over all the datasets of size n. Combining Lemmas SM1.1 and SM1.2, we derive

$$\log \mathcal{N}_{\infty}(u, \mathcal{N}\mathcal{N}(W, L), n) \le C_1 W^2 L^2 \log(W^2 L) \log(u^{-1}n)$$

for a constant C_1 . It follows that

$$\log \mathcal{N}(u, \mathcal{L}_{n,\tau}, L_{\infty}(P_n)) \leq C_2 W^2 L^2 \log(W^2 L) \log(u^{-1} n M_{\tau})$$

for a constant C_2 . Since the class $\mathcal{NN}(W, L, K)$ is bounded and ℓ is continuous, there exists B > 0 such that $\sup_{\boldsymbol{z} \in \mathcal{Z}} |\max_j \ell(f(\boldsymbol{x} + \boldsymbol{\delta}_j), y)| \leq B$ for any $f \in \mathcal{NN}(W, L, K)$. From Lemma SM1.4 and $\log M_{\tau} \leq cd \log(\varepsilon \tau^{-1})$, we have

$$\mathbb{E}_{\sigma} \left\{ \sup_{f \in \mathcal{NN}(W,L,K)} \frac{1}{n} \sum_{i=1}^{n} \sigma_{i} \max_{j} \ell(f(\boldsymbol{x}_{i} + \boldsymbol{\delta}_{j}), y_{i}) \right\}$$

$$\leq \inf_{\delta \geq 0} \left\{ 4\delta + 12 \int_{\delta}^{B} \sqrt{\frac{\log \mathcal{N}(u, \mathcal{L}_{n,\tau}, L_{\infty}(P_{n}))}{n}} du \right\}$$

$$\lesssim \inf_{\delta \geq 0} \left\{ \delta + WL \sqrt{\log(W^{2}L)} n^{-1/2} \cdot \int_{\delta}^{B} \left[\sqrt{\log(u^{-1})} + \sqrt{\log n} + \sqrt{\log M_{\tau}} \right] du \right\}$$

$$\lesssim WL \sqrt{\log(W^{2}L)} n^{-1/2} \left\{ \sqrt{\log n} + \sqrt{\log(\varepsilon\tau^{-1})} \right\}.$$

Therefore, by selecting τ such that $\varepsilon \tau^{-1} = O(n)$, we show

$$I_1 \lesssim K \varepsilon n^{-1} + WL \sqrt{\log(W^2 L)} n^{-1/2} \sqrt{\log n}.$$

Following a similar procedure, we have $I_4 \lesssim K \varepsilon n^{-1} + WL \sqrt{\log(W^2 L)} n^{-1/2} \sqrt{\log n}$.

A.2. Bound for I_2. Define the approximation error by

$$\mathcal{E}\left(\mathcal{H}^{\alpha}, \mathcal{NN}(W, L, K)\right) = \sup_{f \in \mathcal{H}^{\alpha}} \inf_{\phi \in \mathcal{NN}(W, L, K)} \|f - \phi\|_{C([0,1]^d)},$$

where $C([0,1]^d)$ is the space of continuous functions on $[0,1]^d$ equipped with the supremum norm. There exists $\bar{f} \in \mathcal{NN}(W,L,K)$ approximating the target function $f^* \in \mathcal{H}^{\alpha}$ such that

 $\|f^{\star} - \bar{f}\|_{C([0,1]^d)} = O(\mathcal{E}\left(\mathcal{H}^{\alpha}, \mathcal{NN}(W, L, K)\right)).$

The difference between the empirical adversarial risks $\widetilde{\mathcal{R}}_{P_n}(f^\star,\varepsilon)$ and $\widetilde{\mathcal{R}}_{P_n}(\bar{f},\varepsilon)$ satisfies

$$\begin{aligned} \left| \widetilde{\mathcal{R}}_{P_n}(f^{\star},\varepsilon) - \widetilde{\mathcal{R}}_{P_n}(\bar{f},\varepsilon) \right| &\leq \frac{1}{n} \sum_{i=1}^n \left| \sup_{X_i' \in B_{\varepsilon}(X_i)} \ell(f^{\star}(X_i'),Y_i) - \sup_{X_i' \in B_{\varepsilon}(X_i)} \ell(\bar{f}(X_i'),Y_i) \right| \\ &\leq \frac{1}{n} \sum_{i=1}^n \sup_{X_i' \in B_{\varepsilon}(X_i)} \left| \ell(f^{\star}(X_i'),Y_i) - \ell(\bar{f}(X_i'),Y_i) \right| \\ &\leq \operatorname{Lip}^1(\ell) \cdot \| f^{\star} - \bar{f} \|_{C([0,1]^d)}. \end{aligned}$$

Since \widehat{f}_n minimizes the empirical adversarial risk over the class $\mathcal{NN}(W, L, K)$, then

$$I_{2} = \widetilde{\mathcal{R}}_{P_{n}}(\widehat{f}_{n},\varepsilon) - \widetilde{\mathcal{R}}_{P_{n}}(f^{\star},\varepsilon) = \widetilde{\mathcal{R}}_{P_{n}}(\widehat{f}_{n},\varepsilon) - \widetilde{\mathcal{R}}_{P_{n}}(\overline{f},\varepsilon) + \widetilde{\mathcal{R}}_{P_{n}}(\overline{f},\varepsilon) - \widetilde{\mathcal{R}}_{P_{n}}(f^{\star},\varepsilon)$$
$$\leq \widetilde{\mathcal{R}}_{P_{n}}(\overline{f},\varepsilon) - \widetilde{\mathcal{R}}_{P_{n}}(f^{\star},\varepsilon)$$
$$\leq \operatorname{Lip}^{1}(\ell) \cdot \|f^{\star} - \overline{f}\|_{C([0,1]^{d})}.$$

The approximation error $\mathcal{E}(\mathcal{H}^{\alpha}, \mathcal{NN}(W, L, K))$ is investigated by [22] and the result is given as follows.

Lemma A.1 ([22, Theorem 3.2]). Let $\gamma = \lceil \log_2(d+r) \rceil$. There exists c > 0 such that for any $W \ge c(K/\log^{\gamma} K)^{(2d+\alpha)/(2d+2)}$ and $L \ge 4\gamma + 2$,

$$\mathcal{E}(\mathcal{H}^{\alpha}, \mathcal{NN}(W, L, K)) \lesssim (K/\log^{\gamma} K)^{-\alpha/(d+1)}.$$

Therefore, we derive

$$I_2 \lesssim (K/\log^{\gamma} K)^{-\alpha/(d+1)}.$$

A.3. Bound for I₃. For any $f \in \mathcal{H}^{\alpha}$ and $\boldsymbol{z} = (\boldsymbol{x}, y) \in \mathcal{Z}$, we define $\tilde{\ell}(f, \boldsymbol{z}) = \sup_{\boldsymbol{x}' \in B_{\varepsilon}(\boldsymbol{x})} \ell(f(\boldsymbol{x}'), y)$. The error I_3 can be upper bounded by

$$I_3 = \widetilde{\mathcal{R}}_{P_n}(f^\star, \varepsilon) - \widetilde{\mathcal{R}}_P(f^\star, \varepsilon) \le \sup_{f \in \mathcal{H}^\alpha} \{ \mathbb{E}_{Z \sim P_n} \widetilde{\ell}(f, Z) - \mathbb{E}_{Z \sim P} \widetilde{\ell}(f, Z) \}.$$

Define the class $\mathcal{L}^{\alpha} = \{\tilde{\ell}(f, \mathbf{z}) : \mathcal{Z} \mapsto \mathbb{R} \mid f \in \mathcal{H}^{\alpha}\}$. Let $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ consist of i.i.d. Rademacher variables and be independent of the data. For any samples $\mathbf{z}_1, \ldots, \mathbf{z}_n$ from \mathcal{Z} , we denote the empirical Rademacher complexity of the class \mathcal{L}^{α} by

$$\widehat{\Re}_n(\mathcal{L}^{\alpha}) = \mathbb{E}_{\boldsymbol{\sigma}} \left\{ \sup_{f \in \mathcal{H}^{\alpha}} \frac{1}{n} \sum_{i=1}^n \sigma_i \widetilde{\ell}(f, \boldsymbol{z}_i) \right\}.$$

From $\sup_{f \in \mathcal{H}^{\alpha}} \|f\|_{\infty} \leq 1$, there exists a constant B such that $\sup_{z \in \mathcal{Z}} |\tilde{\ell}(f, z)| \leq B$ for any $f \in \mathcal{H}^{\alpha}$. In addition, we have

$$\log \mathcal{N}(u, \mathcal{L}^{\alpha}, \|\cdot\|_{\infty}) \leq \log \mathcal{N}(u/\mathrm{Lip}^{1}(\ell), \mathcal{H}^{\alpha}, \|\cdot\|_{\infty}).$$

This is because for any $f, \tilde{f} \in \mathcal{H}^{\alpha}$ satisfying $\|f - \tilde{f}\|_{\infty} \leq u/\text{Lip}^{1}(\ell)$, it follows that

$$\begin{split} |\tilde{\ell}(f, \boldsymbol{z}) - \tilde{\ell}(\tilde{f}, \boldsymbol{z})| &= \left| \sup_{\boldsymbol{x}' \in B_{\varepsilon}(\boldsymbol{x})} \ell(f(\boldsymbol{x}'), y) - \sup_{\boldsymbol{x}' \in B_{\varepsilon}(\boldsymbol{x})} \ell(\tilde{f}(\boldsymbol{x}'), y) \right| \\ &\leq \operatorname{Lip}^{1}(\ell) \sup_{\boldsymbol{x}' \in B_{\varepsilon}(\boldsymbol{x})} |f(\boldsymbol{x}') - \tilde{f}(\boldsymbol{x}')| \\ &\leq u. \end{split}$$

From [24], $\log \mathcal{N}(u, \mathcal{H}^{\alpha}, \|\cdot\|_{\infty}) \lesssim u^{-d/\alpha}$ holds. Hence,

$$\log \mathcal{N}(u, \mathcal{L}^{\alpha}, L_2(P_n)) \leq \log \mathcal{N}(u, \mathcal{L}^{\alpha}, \|\cdot\|_{\infty}) \lesssim u^{-d/\alpha},$$

where $L_2(P_n)$ denotes the L_2 metric generated by the samples. It follows by Lemma SM1.4 that

$$\widehat{\Re}_n(\mathcal{L}^{\alpha}) \lesssim \inf_{\delta \ge 0} \left\{ 4\delta + 12 \int_{\delta}^1 \sqrt{\frac{\log \mathcal{N}(u, \mathcal{L}^{\alpha}, \|\cdot\|_{\infty})}{n}} du \right\}.$$

Let $\gamma = d/(2\alpha)$. Thus, we show

$$\widehat{\Re}_n(\mathcal{L}^{\alpha}) \lesssim \inf_{\delta \ge 0} \left(\delta + n^{-1/2} \int_{\delta}^1 u^{-\gamma} du \right).$$

BOUNDS FOR ADVERSARIAL EXCESS RISK

When $\gamma > 1$, by taking $\delta = n^{-\alpha/d}$, one has

$$\widehat{\Re}_n(\mathcal{L}^{\alpha}) \lesssim \inf_{\delta \ge 0} \left(\delta + (\gamma - 1)^{-1} n^{-1/2} (\delta^{1 - \gamma} - 1) \right) \lesssim n^{-\alpha/d}.$$

When $\gamma = 1$, by taking $\delta = n^{-1/2}$, one has

$$\widehat{\Re}_n(\mathcal{L}^{\alpha}) \lesssim \inf_{\delta \ge 0} \left(\delta - n^{-1/2} \log \delta \right) \lesssim n^{-1/2} \log n.$$

When $\gamma < 1$, one has

$$\widehat{\Re}_n(\mathcal{L}^{\alpha}) \lesssim \inf_{\delta \ge 0} \left(\delta + (1-\gamma)^{-1} n^{-1/2} (1-\delta^{1-\gamma}) \right) \lesssim n^{-1/2}.$$

Combining these cases together, we derive

$$\widehat{\Re}_n(\mathcal{L}^{\alpha}) \lesssim n^{-\min\{1/2,\alpha/d\}} \log^{c(\alpha,d)} n,$$

where $c(\alpha, d) = 1$ if $d = 2\alpha$, and $c(\alpha, d) = 0$ otherwise. Following a similar analysis method in (A.1), we derive

$$\sup_{f \in \mathcal{H}^{\alpha}} \left\{ \mathbb{E}_{Z \sim P_n} \tilde{\ell}(f, Z) - \mathbb{E}_{Z \sim P} \tilde{\ell}(f, Z) \right\} \lesssim n^{-\min\{1/2, \alpha/d\}} \log^{c(\alpha, d)} n.$$

Consequently, it follows that

$$I_3 \leq n^{-\min\{1/2,\alpha/d\}} \log^{c(\alpha,d)} n.$$

Let $\gamma = \lceil \log_2(d+r) \rceil$. Combining the results from sections A.1–A.3, we show for any $W \ge c(K/\log^{\gamma} K)^{(2d+\alpha)/(2d+2)}$ and $L \ge 4\gamma + 2$,

(A.2)

$$\begin{aligned}
\mathcal{E}(f_n,\varepsilon) &= \max\{\mathcal{E}_1, \mathcal{E}_2\} \\
&\leq \max\{I_1 + I_2 + I_3, I_1 + I_4\} \\
&\lesssim K\varepsilon n^{-1} + WL\sqrt{\log(W^2L)}n^{-1/2}\sqrt{\log n} \\
&+ (K/\log^{\gamma}K)^{-\alpha/(d+1)} + n^{-\min\{1/2,\alpha/d\}}\log^{c(\alpha,d)}n.
\end{aligned}$$

where $c(\alpha, d) = 1$ when $d = 2\alpha$, and $c(\alpha, d) = 0$ otherwise. By selecting $K \simeq n^{(d+1)/(2d+3\alpha)}$, and $WL \simeq n^{(2d+\alpha)/(4d+6\alpha)}$, we have

$$WL\sqrt{\log(W^{2}L)}n^{-1/2}\sqrt{\log n} + (K/\log^{\gamma}K)^{-\alpha/(d+1)} \\ \lesssim n^{-\alpha/(2d+3\alpha)}\log n^{\max\{1,\gamma\alpha/(d+1)\}}.$$

This leads to

$$\widetilde{\mathcal{E}}(\widehat{f}_n,\varepsilon) \lesssim K \varepsilon n^{-1} + n^{-\alpha/(2d+3\alpha)} \log n^{\xi}$$

where $\xi = \max\{1, \gamma \alpha/(d+1)\}$. Hence, the result is proved.

Acknowledgments. We are grateful to the AE and two anonymous reviewers for their constructive comments that helped improve the quality of the paper.

REFERENCES

- A. ATHALYE, N. CARLINI, AND D. WAGNER, Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, in Proceedings of the 35th International Conference on Machine Learning, Vol. 80, PMLR, 2018, pp. 274–283.
- [2] P. AWASTHI, N. FRANK, A. MAO, M. MOHRI, AND Y. ZHONG, Calibration and consistency of adversarial surrogate losses, in Advances in Neural Information Processing Systems, Vol. 34, Curran Associates, 2021, pp. 9804–9815.
- [3] P. AWASTHI, N. FRANK, AND M. MOHRI, Adversarial learning guarantees for linear hypotheses and neural networks, in Proceedings of the 37th International Conference on Machine Learning, Vol. 119, PMLR, 2020, pp. 431–441.
- [4] P. AWASTHI, N. FRANK, AND M. MOHRI, On the existence of the adversarial Bayes classifier, in Advances in Neural Information Processing Systems, Vol. 34, Curran Associates, 2021, pp. 2978–2990.
- [5] P. AWASTHI, A. MAO, M. MOHRI, AND Y. ZHONG, A finer calibration analysis for adversarial robustness, in Proceedings of the 40th International Conference on Machine Learning, Vol. 202, PMLR, 2023, pp. 1373–1391.
- [6] H. BAO, C. SCOTT, AND M. SUGIYAMA, Calibrated surrogate losses for adversarially robust classification, in Proceedings of the 33rd Conference on Learning Theory, Vol. 125, PMLR, 2020, pp. 408–451.
- B. BAUER AND M. KOHLER, On deep learning as a remedy for the curse of dimensionality in nonparametric regression, Ann. Statist., 47 (2019), pp. 2261–2285.
- [8] W. BRENDEL, J. RAUBER, AND M. BETHGE, *Decision-based adversarial attacks: Reliable attacks against black-box machine learning models*, in International Conference on Learning Representations, 2018.
- S. BUBECK AND M. SELLKE, A universal law of robustness via isoperimetry, in Advances in Neural Information Processing Systems, Vol. 34, Curran Associates, 2021, pp. 28811–28822.
- [10] L. BUNGERT, N. GARCÍA TRILLOS, AND R. MURRAY, The geometry of adversarial training in binary classification, Inf. Inference, 12 (2023), pp. 921–968.
- [11] N. CARLINI AND D. WAGNER, Towards evaluating the robustness of neural networks, in 2017 IEEE Symposium on Security and Privacy (SP), IEEE Computer Society, Los Alamitos, CA, 2017, pp. 39–57.
- [12] M. CISSE, P. BOJANOWSKI, E. GRAVE, Y. DAUPHIN, AND N. USUNIER, Parseval networks: Improving robustness to adversarial examples, in Proceedings of the 34th International Conference on Machine Learning, Vol. 70, PMLR, 2017, pp. 854–863.
- [13] J. COHEN, E. ROSENFELD, AND Z. KOLTER, Certified adversarial robustness via randomized smoothing, in Proceedings of the 36th International Conference on Machine Learning, Vol. 97, PMLR, 2019, pp. 1310–1320.
- [14] C. DAN, Y. WEI, AND P. RAVIKUMAR, Sharp statistical guarantees for adversarially robust Gaussian classification, in Proceedings of the 37th International Conference on Machine Learning, Vol. 119, PMLR, 2020, pp. 2345–2355.
- [15] E. DOBRIBAN, H. HASSANI, D. HONG, AND A. ROBEY, Provable tradeoffs in adversarially robust classification, in Proceedings of the 37th Conference on Machine Learning, Vol. 119, PMLR, 2020, pp. 2595–2605.
- [16] N. S. FRANK, Existence and minimax theorems for adversarial surrogate risks in binary classification, J. Mach. Learn. Res., 25 (2024), 58.
- [17] I. J. GOODFELLOW, J. SHLENS, AND C. SZEGEDY, Explaining and harnessing adversarial examples, in 3rd International Conference on Learning Representations, ICLR 2015.
- [18] M. HEIN AND M. ANDRIUSHCHENKO, Formal guarantees on the robustness of a classifier against adversarial manipulation, in Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, 2017, pp. 2263–2273.
- [19] A. JAVANMARD AND M. SOLTANOLKOTABI, Precise statistical analysis of classification accuracies for adversarial training, Ann. Statist., 50 (2022), pp. 2127–2156.

- [20] A. JAVANMARD, M. SOLTANOLKOTABI, AND H. HASSANI, Precise tradeoffs in adversarial training for linear regression, in Proceedings of the 33rd Conference on Learning Theory, Vol. 125, PMLR, 2020, pp. 2034–2078.
- [21] Y. JIAO, G. SHEN, Y. LIN, AND J. HUANG, Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors, Ann. Statist., 51 (2023), pp. 691–716.
- [22] Y. JIAO, Y. WANG, AND Y. YANG, Approximation bounds for norm constrained neural networks with applications to regression and GANs, Appl. Comput. Harmon. Anal., 65 (2023), pp. 249–278.
- [23] J. KHIM AND P.-L. LOH, Adversarial risk bounds via function transformation, in Proceedings of the 35th International Conference on Machine Learning, Vol. 80, PMLR, 2018, pp. 2621–2630.
- [24] A. N. KOLMOGOROV AND V. M. TIKHOMIROV, ε-entropy and ε-capacity of sets in function spaces, Uspekhi Mat. Nauk, 14 (1959), pp. 3–86.
- [25] J. LEE AND M. RAGINSKY, Minimax statistical learning with Wasserstein distances, in Advances in Neural Information Processing Systems, Vol. 31, Curran Associates, 2018, pp. 2692–2701.
- [26] J. LU, Z. SHEN, H. YANG, AND S. ZHANG, Deep network approximation for smooth functions, SIAM J. Math. Anal., 53 (2021), pp. 5465–5506, https://doi.org/10.1137/20M134695X.
- [27] X. MA, Z. WANG, AND W. LIU, On the tradeoff between robustness and fairness, in Advances in Neural Information Processing Systems, Vol. 34, Curran Associates, 2021, pp. 26230–26241.
- [28] A. MADRY, A. MAKELOV, L. SCHMIDT, D. TSIPRAS, AND A. VLADU, Towards deep learning models resistant to adversarial attacks, in International Conference on Learning Representations, 2018.
- [29] L. MEUNIER, R. ETTEDGUI, R. PINOT, Y. CHEVALEYRE, AND J. ATIF, Towards consistency in adversarial classification, in Advances in Neural Information Processing Systems, Vol. 35, Curran Associates, 2022, pp. 29947–29959.
- [30] S.-M. MOOSAVI-DEZFOOLI, A. FAWZI, AND P. FROSSARD, DeepFool: A simple and accurate method to fool deep neural networks, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 2574–2582.
- [31] W. MUSTAFA, Y. LEI, AND M. KLOFT, On the generalization analysis of adversarial learning, in Proceedings of the 39th International Conference on Machine Learning, Vol. 162, PMLR, 2022, pp. 16174–16196.
- [32] N. PAPERNOT, P. MCDANIEL, S. JHA, M. FREDRIKSON, Z. B. CELIK, AND A. SWAMI, *The limitations of deep learning in adversarial settings*, in 2016 IEEE European Symposium on Security and Privacy (EuroS&P), IEEE, 2016, pp. 372–387.
- [33] N. PAPERNOT, P. MCDANIEL, X. WU, S. JHA, AND A. SWAMI, Distillation as a defense to adversarial perturbations against deep neural networks, in 2016 IEEE Symposium on Security and Privacy (SP), IEEE, 2016, pp. 582–597.
- [34] M. S. PYDI AND V. JOG, Adversarial risk via optimal transport and optimal couplings, IEEE Trans. Inform. Theory, 67 (2021), pp. 6031–6052.
- [35] M. S. PYDI AND V. JOG, The many faces of adversarial risk, in Advances in Neural Information Processing Systems, Vol. 34, Curran Associates, 2021, pp. 10000–10012.
- [36] A. RAGHUNATHAN, J. STEINHARDT, AND P. LIANG, Certified defenses against adversarial examples, in International Conference on Learning Representations, 2018.
- [37] A. RAGHUNATHAN, S. M. XIE, F. YANG, J. DUCHI, AND P. LIANG, Adversarial training can hurt generalization, in ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena, 2019.
- [38] L. SCHMIDT, S. SANTURKAR, D. TSIPRAS, K. TALWAR, AND A. MADRY, Adversarially robust generalization requires more data, in Advances in Neural Information Processing Systems, Vol. 31, Curran Associates, 2018, pp. 5026–5041.
- [39] J. SCHMIDT-HIEBER, Nonparametric regression using deep neural networks with ReLU activation function, Ann. Statist., 48 (2020), pp. 1875–1897.
- [40] I. STEINWART AND A. CHRISTMANN, Support Vector Machines, Inf. Sci. Stat., Springer, New York, 2008.
- [41] C. J. STONE, Optimal global rates of convergence for nonparametric regression, Ann. Statist., 10 (1982), pp. 1040–1053.
- [42] J. SU, D. V. VARGAS, AND K. SAKURAI, One pixel attack for fooling deep neural networks, IEEE Trans. Evol. Comput., 23 (2019), pp. 828–841.

- [43] C. SZEGEDY, W. ZAREMBA, I. SUTSKEVER, J. BRUNA, D. ERHAN, I. J. GOODFELLOW, AND R. FERGUS, *Intriguing properties of neural networks*, in 2nd International Conference on Learning Representations, ICLR 2014.
- [44] D. TSIPRAS, S. SANTURKAR, L. ENGSTROM, A. TURNER, AND A. MADRY, Robustness may be at odds with accuracy, in International Conference on Learning Representations, 2019.
- [45] Y. TSUZUKU, I. SATO, AND M. SUGIYAMA, Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks, Adv. Neural Inf. Process. Syst., 31 (2018).
- [46] Z. TU, J. ZHANG, AND D. TAO, Theoretical analysis of adversarial learning: A minimax approach, in Advances in Neural Information Processing Systems, Vol. 32, Curran Associates, 2019, pp. 12280– 12290.
- [47] R. VERSHYNIN, High-Dimensional Probability: An Introduction with Applications in Data Science, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2018.
- [48] T.-W. WENG, H. ZHANG, P.-Y. CHEN, J. YI, D. SU, Y. GAO, C.-J. HSIEH, AND L. DANIEL, Evaluating the robustness of neural networks: An extreme value theory approach, in International Conference on Learning Representations, 2018.
- [49] Y. XING, R. ZHANG, AND G. CHENG, Adversarially robust estimate and risk analysis in linear regression, in Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, Vol. 130, PMLR, 2021, pp. 514–522.
- [50] Y.-Y. YANG, C. RASHTCHIAN, H. ZHANG, R. R. SALAKHUTDINOV, AND K. CHAUDHURI, A closer look at accuracy vs. robustness, in Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, 2020, pp. 8588–8601.
- [51] D. YAROTSKY, Optimal approximation of continuous functions by very deep ReLU networks, in Proceedings of the 31st Conference On Learning Theory, Vol. 75, PMLR, 2018, pp. 639–649.
- [52] D. YIN, R. KANNAN, AND P. BARTLETT, Rademacher complexity for adversarially robust generalization, in Proceedings of the 36th International Conference on Machine Learning, Vol. 97, PMLR, 2019, pp. 7085–7094.
- [53] H. ZHANG, Y. YU, J. JIAO, E. XING, L. E. GHAOUI, AND M. JORDAN, Theoretically principled trade-off between robustness and accuracy, in Proceedings of the 36th International Conference on Machine Learning, Vol. 97, PMLR, 2019, pp. 7472–7482.