On the generalization analysis of adversarial learning

Mustafa, Waleed; Lei, Yunwen; Kloft, Marius

On the Generalization Analysis of Adversarial Learning

Waleed Mustafa [1] **Yunwen Lei** [2] **Marius Kloft** [1]

Abstract

Many recent studies have highlighted the susceptibility of virtually all machine-learning models to adversarial attacks. Adversarial attacks are imperceptible changes to an input example of a given prediction model. Such changes are carefully designed to alter the otherwise correct prediction of the model. In this paper, we study the generalization properties of adversarial learning. In particular, we derive high-probability generalization bounds on the adversarial risk in terms of the empirical adversarial risk, the complexity of the function class, and the adversarial noise set. Our bounds are generally applicable to many models, losses, and adversaries. We showcase its applicability by deriving adversarial generalization bounds for the multi-class classification setting and various prediction models (including linear models and Deep Neural Networks). We also derive optimistic adversarial generalization bounds for the case of smooth losses. These are the first fast-rate bounds valid for adversarial deep learning to the best of our knowledge.

1. Introduction

Machine learning has been shown to be susceptible to a large number of security threats (Barreno et al., 2010; Papernot et al., 2016). One such threat is *adversarial examples* (Szegedy et al., 2013; Biggio et al., 2013). Adversarial examples are perturbed inputs carefully designed to alter a model's prediction while being imperceptible to humans. This paper studies the generalization properties of models trained to withstand such adversarial attacks. While much work has been conducted on the algorithmic design of attacks (Carlini et al., 2017; Brendel et al., 2017; Awasthi et al., 2021; Engstrom et al., 2019), and defenses (Madry et al., 2017; Finlay & Oberman, 2019; Kannan et al., 2018), there is a lack of theoretical understanding of the generalization properties of adversarial learning.

The current state of the art in the generalization analysis of adversarial learning has the following limitations to our knowledge. First, all previous papers on adversarial generalization apply solely to *additive* adversaries (in which the attacker *adds* a small perturbation to the input sample). However, there is a recent trend in using *non*-additive adversaries (Hendrycks & Dietterich, 2018; Engstrom et al., 2019; Wong & Kolter, 2020; Awasthi et al., 2021). Previous generalization analyses are inapplicable in this setting. Second, all previous papers consider restricted setups based on simple models and architectures (linear model or one-hidden-layer neural networks; Yin et al., 2019; Awasthi et al., 2020) or unrealistic assumptions (Dan et al., 2020; Schmidt et al., 2018; Farnia et al., 2018; Gao & Wang, 2021), or they use surrogate losses (Khim & Loh, 2018; Yin et al., 2019). However, a surrogate upper bound does not necessarily lead to a meaningful generalization bound on the original (adversarial) loss (Awasthi et al., 2020). Next, there exists no unified analysis applying to a large variety of models, Moreover, there are no fast-rate bounds (Srebro et al., 2010) for adversarial learning using neural networks. Finally, all previous results scale at least $O(\sqrt{K})$ in the number of label classes K and are thus inapplicable to extreme classification, or structured prediction (Prabhu & Varma, 2014).

We derive generalization bounds that do not suffer from the limitations mentioned above. Our contributions can be summarized as follows:

- We derive the first generalization bounds for adversarial learning valid for general (possibly non-additive) noise functions (thus covering a wide array of attacks).
- Our bounds are modular and general. They can be applied to many models, from linear models over kernel machines to neural networks. Extending it to new models is easy: it requires simply computing the ℓ_∞-covering number of the model.
- We show the first generalization bounds for adversarial learning of Deep Neural Networks applying to the adversarial loss directly, not a surrogate.

[1]Department of Computer Science, University of Kaiserslautern, Germany [2]School of Computer Science, University of Birmingham, United Kingdom. Correspondence to: Yunwen Lei <y.lei@bham.ac.uk>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

- Our bounds scale $O(\log(K))$ in the number of label classes, making them suitable for multi-label learning and structured prediction in adversarial environments.
- Finally, we show the first fast-rate bounds for adversarial deep learning.

The rest of the paper is organized as follows. We discuss related work in Section 2. In Section 3, we introduce the notation and the problem setup. Section 4 is dedicated to the main results of this paper. We apply our approach to several models and adversarial attacks in Section 5. Finally, we provide fast-rate bounds in Section 6 and discuss our findings in Section 7.

2. Related Work

In this section, we first give an overview of popular adversarial attacks and defenses and then review the related work on the generalization analysis of adversarial learning.

Adversarial attacks Adversarial attacks are usually categorized as *white-box* (Carlini & Wagner, 2017) or *black-box* (Brendel et al., 2017), depending on the information available to the attacker. Most commonly, the attacker is constrained to alter the input by additive noise from an ℓ_p-ball. Recently, further (non-additive) attack models have been considered. In which the adversary manipulates the input by a non-linear transformation, either in the input space (e.g., rotation of an input image; Engstrom et al., 2019) or in a semantic representation space (e.g., in the frequency domain of an image; Awasthi et al., 2021).

Defenses In response to such attacks, several defense mechanisms have been developed, for instance, based on regularizing the model's Lipschitz constant (Bietti et al., 2018; Cissé et al., 2017), input gradient (Hein & Andriushchenko, 2017; Ross & Doshi-Velez, 2018), or input Hessian (Mustafa et al., 2020) at training. The most widely used defense mechanism against adversarial attacks is *adversarial training* (Madry et al., 2017) and its variants (Kannan et al., 2018; Zhang et al., 2019). Its key idea is to replace clean training samples with their adversarial counterparts while maintaining their correct labels. Systematic studies have shown that the resulting models are robust and can withstand a large number of attacks (Athalye et al., 2018).

Generalization Analysis of Adversarial Learning We now discuss the existing generalization bounds for adversarial learning. Dan et al. (2020) and Schmidt et al. (2018) showed bounds valid only in the idealized binary classification scenario where the data is sampled from two Gaussians. The prediction function is linear in both papers, and the bounds scale linearly in the number of dimensions. Dan et al. (2020) also showed a matching lower bound. Attias et al. (2019) showed a bound based on the VC-dimension of the function class, which can be very large for many models. Their study considers attacks with a finite adversarial noise set. However, virtually all practical attacks use an uncountable infinite noise set. Gao & Wang (2021) and Farnia et al. (2018) showed bounds assuming that the attack is apriori known to the learner, which is a strong assumption since the attacker could utilize any attack available in practice. Xing et al. (2021) leveraged the algorithmic stability to study the generalization of adversarial learning.

Several authors have used the Rademacher complexity to study the generalization of ℓ_p-additive-perturbation attacks (Khim & Loh, 2018; Yin et al., 2019; Awasthi et al., 2020; Xiao et al., 2021). In contrast, our analysis applies to a much wider range of attacks. Khim & Loh (2018) introduced the *tree-transform*, in which the supremum over the adversarial noise set is propagated through the network layers to establish an upper-bound on the adversarial loss. This upper bound, however, can be vacuous for Deep Neural Networks since its looseness grows exponentially with the depth of the network. Since our approach applies directly to the loss, it does not suffer from this problem. In addition, their bound grows as $O(K)$ in the number of classes K, while ours is $O(\log(K))$.

Yin et al. (2019) showed generalization bounds for linear models and one-hidden-layer neural networks based on the surrogate loss introduced in Raghunathan et al. (2018) under ℓ_∞-additive perturbation. Their bound does not apply to (deep) neural networks with two or more hidden layers. Our approach applies to neural networks with arbitrary many layers (Deep Neural Networks) and is directly based on the adversarial loss, not on a surrogate. Awasthi et al. (2020) also derived bounds directly based on the adversarial loss, but only for linear models and one-hidden-layer networks. Their bound scales as $O(\sqrt{m})$ in the number of neurons m, while ours is $O(\log(m))$.

3. Problem Setup

We now define the formal setup of adversarial learning. We start by defining general statistical learning, after which we introduce the adversary. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ be an input-output space, where $\mathcal{X} \subset \mathbb{R}^d$ is an input/feature space and $\mathcal{Y} = [K] := \{1, \dots, K\}$ is an output/label space. We further assume that there is an unknown probability measure $\mathcal{D}$ defined over $\mathcal{Z}$. The goal of supervised learning is to find a function $g : \mathcal{X} \to \mathcal{Y}$, such that, for a given loss function $\ell_c : \mathcal{Y} \times \mathcal{Y}$, the expected loss $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_c(g(x), y)]$ is minimized. We are interested in parameterized-scoring-based classifiers based on multivariate model functions $f : \mathcal{X} \to \mathbb{R}^K$. The classification function g is obtained by $g(x) = \arg\max_{k\in\mathcal{Y}} f(x)_k$. Since the distribution $\mathcal{D}$ is

usually unknown, we utilize a sample from it to learn g. Let $\{z_i = (x_i, y_i)\}_{i=1}^n$ be an i.i.d. sample drawn from $\mathcal{D}$. The classification function is selected from the hypothesis class $\mathcal{F}_{\mathbb{W}} := \{x \mapsto f(x, w) : x \in \mathcal{X}, w \in \mathbb{W}\}$ parameterized by $w \in \mathbb{W}$, where $\mathbb{W}$ is some parameter space. We are interested in empirical risk minimization, with the parameter $\hat{w}$ defined by

$$\hat{w} = \arg\min_{w \in \mathbb{W}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i, w), y_i),$$

where $\ell : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}_+$ is a given loss function.

An adversary utilizes a noise application function $A : \mathcal{X} \times \mathcal{B} \to \mathcal{X}$, where $\mathcal{B}$ is a noise space, to modify an input of a classifier with the goal to alter its prediction. To our knowledge, all previous work on adversarial learning theory considered only the special case where A is the additive noise $A(x, \delta) = x + \delta$ and $\mathcal{B}$ is the ℓ_p-ball $\{\delta : \|\delta\|_p \leq \epsilon\}$. In contrast, we consider a more general A[1] and therefore, our results can be applied to a broader array of attacks (e.g., Engstrom et al., 2019; Awasthi et al., 2021). Given an input example x and a learned parameter setting w, the adversary selects the noise parameter δ^* by

$$\delta^* = \arg\max_{\delta \in \mathcal{B}} \ell(f(A(x, \delta); w), y).$$

A common strategy to train a robust model is adversarial training (Madry et al., 2017). The training is achieved by solving the min-max problem

$$\hat{w}_{\text{adv}} = \arg\min_{w \in \mathbb{W}} \hat{R}_{\text{adv}}(w),$$

where $\hat{R}_{\text{adv}}(w) := \frac{1}{n} \sum_{i=1}^{n} \max_{\delta \in \mathcal{B}} \ell(f(A(x_i, \delta); w), y_i)$ measures the empirical risk of the model on the training examples subject to adversarial noise. We are interested in the adversarial generalization behavior of $\hat{w}_{\text{adv}}$ as measured by the population risk in the adversarial setting:

$$R_{\text{adv}}(\hat{w}_{\text{adv}}) := \mathbb{E}_{\mathcal{D}}\big[\max_{\delta \in \mathcal{B}} \ell(f(A(x, \delta)\ \hat{w}_{\text{adv}}))\big].$$

We refer to the difference between the population risk $R_{\text{adv}}(w)$ and the empirical risk $\hat{R}_{\text{adv}}(w)$ as the generalization error of w.

4. Main Results

In this section, we introduce our main result. Our primary tool is based on *covering numbers* defined below. Roughly speaking, covering numbers measure the complexity of a function class $\mathcal{F}$ in terms of the number of balls required to approximate the class to a prescribed accuracy under the metric D.

Definition 4.1 (Covering number)**.** Let $(\mathcal{A}, D)$ be a pseudometric space. We say that $C \subset \mathcal{A}$ is an (ϵ, D)-cover to $A \subset \mathcal{A}$ if

$$\sup_{a \in A} \inf_{c \in C} D(a, c) \leq \epsilon.$$

The covering number of A at ϵ precision, denoted as $\mathcal{N}(\epsilon, A)$, is the size of the minimal-cardinality set that covers A.

For a data set $S := \{z_i\}_{i=1}^n$ with $z_i \in \mathcal{Z}$ and a function class $\mathcal{F}$ with its elements taking values in a (possibly infinite-dimensional) real vector space $\mathcal{V}$, the (ϵ, D)-empirical covering number of $\mathcal{F}$ on S, denoted as $\mathcal{N}_D(\epsilon, \mathcal{F}, S)$, is the (ϵ, D)-covering number of the set

$$\mathcal{F}|_S = \{(f(z_1), \ldots, f(z_n)) : f \in \mathcal{F}\}.$$

The ℓ_p pseudo-metric on $\mathcal{V}^n$ is

$$D_p(x, y) = (\frac{1}{n} \sum_{i=1}^{n} \|x_i - y_i\|^p)^{\frac{1}{p}},$$

where $\|\cdot\|$ is a general norm on $\mathcal{V}$. If $p = \infty$ we obtain $D_\infty(x, y) := \max_{i \in [n]} \|x_i - y_i\|$. We denote by $\mathcal{N}_p(\epsilon, \mathcal{F}, \|\cdot\|, S)$ and $\mathcal{N}_\infty(\epsilon, \mathcal{F}, \|\cdot\|, S)$ the covering numbers w.r.t. the D_p and D_∞ metrics, respectively.

Our main approach is to view adversarial generalization as multi-label classification, which allows us to utilize recent advances in the generalization analysis of vector-valued learning to study adversarial generalization (Wu et al., 2021). The loss function class of interest is

$$\mathcal{G}_{\text{adv}} := \{z \mapsto \max_{\delta \in \mathcal{B}} \ell(f(A(x, \delta)), y) : f \in \mathcal{F}\}. \quad (1)$$

We define a new function class as follows. For each function $f \in \mathcal{F}$, we construct a functional $g : \mathcal{Z} \to (\mathbb{R}^K)^{\mathcal{B}}$ as $g(z) := \ell(f(A(x, \cdot)), y)$. That is, g receives an input-output vector z and outputs a function $\mathcal{B} \mapsto \mathbb{R}^K$. The corresponding function class is then

$$\mathcal{G} := \{z \mapsto \ell(f(A(x, \cdot)), y) : f \in \mathcal{F}\}.$$

In the following lemma, we introduce our first main result. The lemma states that the covering number of the function class $\mathcal{G}_{\text{adv}}$ is bounded by the covering number of the function class $\mathcal{G}$. Therefore, the covering number of $\mathcal{G}$ can be used to control the adversarial generalization.

Lemma 4.2. *Let $\mathcal{G}_{\text{adv}}$ and $\mathcal{G}$ be defined as above. It holds*

$$\mathcal{N}_\infty(\epsilon, \mathcal{G}_{\text{adv}}, |\cdot|, S) \leq \mathcal{N}_\infty(\epsilon, \mathcal{G}, \|\cdot\|_\infty, S),$$

where $\|g\|_\infty = \sup_z |g(z)|, g \in \mathcal{G}$.

[1] Since any sample x' can be obtained from another sample x by adding $\delta = x' - x$, the key limitation of additive attacks lies in the restriction $\|\delta\|_p \leq \epsilon$. For example, while a rotation of an image by a small angle is considered imperceptible, it may result in $\|x' - x\|_\infty \leq 1$, where pixels are in $[0, 1]$.

The proof of this lemma is deferred to the appendix. The theorem allows us to control the ℓ_∞-covering number of the adversarial loss class by the ℓ_∞-covering number of the class $\mathcal{G}$.

While Lemma 4.2 provides an essential tool to control the complexity of the adversarial loss class, deriving an upper bound to $\mathcal{N}_\infty(\epsilon, \mathcal{G}, \|.\|_\infty, S)$ is not simple. The main challenge is that the functions in $\mathcal{G}$ take values in an infinite-dimensional vector space. Our approach to this problem is to approximate such a space by a finite discretization. To that end, we introduce a Lipschitzness assumption on the functions $\delta \mapsto \ell(f(A(x,\delta)), y)$ for $f \in \mathcal{F}$.

Definition 4.3 (Lipschitz continuity)**.** Let $f : \mathcal{V}_1 \to \mathcal{V}_2$ be a map from vector space $\mathcal{V}_1$ to $\mathcal{V}_2$. Let $\mathcal{V}_1$ and $\mathcal{V}_2$ be endowed with norms $\|\cdot\|_{l_1}$ and $\|\cdot\|_{l_2}$, respectively. We say that f is $(\|\cdot\|_{l_2}, \|\cdot\|_{l_1})$-Lipschitz with constant τ if, for any $\delta, \tilde{\delta} \in \mathcal{V}_1$, we have:

$$\|f(\delta) - f(\delta')\|_{l_2} \leq \tau \|\delta - \tilde{\delta}\|_{l_1}.$$

When $\mathcal{V}_2 = \mathbb{R}$, then $\|\cdot\|_{l_2}$ is the absolute value and the notation is simplified to $\|\cdot\|_{l_1}$-Lipschitz.

We now introduce our main result to relate the ℓ_∞-covering number of class $\mathcal{G}_{\text{adv}}$ to the covering number of the discretized version $\tilde{\mathcal{G}}_{\text{adv}}$ defined below.

Lemma 4.4. *Let $\delta \mapsto \ell(f(A(x,\delta)), y)$ be $\|\cdot\|$-Lipschitz with constant L, for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $f \in \mathcal{F}$. Let $C_{\mathcal{B}}(\epsilon/2L)$ be an $(\epsilon/2L, \|\cdot\|)$-cover of $\mathcal{B}$. We define the loss class*

$$\tilde{\mathcal{G}}_{\text{adv}} = \left\{(z,\delta) \mapsto \ell(f(A(x,\delta)), y) : f \in \mathcal{F}\right\} \tag{2}$$

and the extended training set

$$\tilde{S} = \{(x_i, \tilde{\delta}, y_i) : i \in [n], \tilde{\delta} \in C_{\mathcal{B}}(\epsilon/2L)\}. \tag{3}$$

Then we have

$$\begin{aligned}\mathcal{N}_\infty(\epsilon, \mathcal{G}_{\text{adv}}, |\cdot|, S) &\leq \mathcal{N}_\infty(\epsilon, \mathcal{G}, \|\cdot\|_\infty, S) \\ &\leq \mathcal{N}_\infty(\epsilon/2, \tilde{\mathcal{G}}_{\text{adv}}, |\cdot|, \tilde{S}).\end{aligned}$$

Note that functions in $\mathcal{G}_{\text{adv}}$ involve the maximum over $\mathcal{B}$ due to adversarial learning, which is removed for functions in $\tilde{\mathcal{G}}_{\text{adv}}$. This is achieved by incorporating δ into the argument. Also, note that each element of $\mathcal{G}_{\text{adv}}$ and $\tilde{\mathcal{G}}_{\text{adv}}$ is a scalar function. Therefore we use $|\cdot|$ in the definition of covering numbers. For comparison, each element in $\mathcal{G}$ is a functional mapping z to a function on $\mathcal{B}$. Therefore we use $\|\cdot\|_\infty$ in the involved covering number. For brevity of notation, we often omit either $|\cdot|$ or $\|\cdot\|_\infty$ when mentioning covering numbers. The proof of this lemma is deferred to the appendix.

Remark 4.5. We note that Lemma 4.4 controls the complexity of the adversarial loss class by the complexity of its non-adversarial counterpart. Therefore, it can be readily applied to a wide array of models where a covering number bound exists. This is in contrast to virtually all approaches in the literature (Yin et al., 2019; Khim & Loh, 2018; Xiao et al., 2021; Awasthi et al., 2020; Farnia et al., 2018), in which a function-class-specific approach is used.

Remark 4.6. The Lipschitzness condition on the function $\delta \mapsto \ell(f(A(x,\delta)), y)$ is necessary in Lemma 4.4. It is a mild condition that most attacks fulfill (e.g., Engstrom et al., 2019; Awasthi et al., 2021; Madry et al., 2017).

Remark 4.7. The size of the extended training set $\tilde{S}$ grows linearly in the size of the set $C_{\mathcal{B}}(\epsilon/2L)$. In principle, the size $C_{\mathcal{B}}(\epsilon/2L)$ can grow exponentially in the dimensionality of $\mathcal{B}$. However, the dependence of the generalization performance is of the order $O(\log^{\frac{1}{2}}(|\tilde{S}|))$ for many function classes (e.g., Bartlett et al., 2017; Zhang, 2002; Mustafa et al., 2021). These lead to generalization bounds with a square-root dependency on the dimensionality of $\mathcal{B}$.

We now state our main generalization bound, which controls the generalization errors by an integral of covering numbers on $\tilde{\mathcal{G}}_{\text{adv}}$. This result is modular: one needs to plug a covering-number bound into it to obtain a generalization bound for adversarial learning.

Theorem 4.8. *Let $\delta \in (0,1)$. Suppose that the loss ℓ is bounded by 1. Let $\tilde{\mathcal{G}}$ and $\tilde{S}$ be defined as (2) and (3). With probability at least $1-\delta$ over the training data S, for all $w \in \mathbb{W}$, we have*

$$\begin{aligned}R_{\text{adv}}(w) \leq \hat{R}_{\text{adv}}(w) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \\ + \inf_{\alpha>0}\left(8\alpha + \frac{24}{\sqrt{n}}\int_\alpha^1 \sqrt{\log \mathcal{N}_\infty(\epsilon/2, \tilde{\mathcal{G}}_{\text{adv}}, \tilde{S})}d\epsilon\right).\end{aligned}$$

5. Applications

In this section, we present several applications of Theorem 4.8. Our first example is multi-class linear classifiers with additive adversary.

5.1. Regularized Multi-class Linear Classifiers

We consider the K-class linear classifiers with the following hypothesis class:

$$\mathcal{F} := \left\{x \mapsto Wx : W \in \mathbb{R}^{K\times d}, \|W\|_{2,2} \leq \Lambda\right\}. \tag{4}$$

We further assume that $\max_{x\in\mathcal{X}} \|x\| \leq \Psi$. For a given hypothesis $f \in \mathcal{F}$, we carry out prediction for an input $x \in \mathcal{X}$ by $x \mapsto \arg\max_{i\in[K]} f_i(x)$. The quality of prediction is measured by the multi-class margin loss defined as

$$\ell_\rho(t,y) = \begin{cases}1, & \text{if } M(t,y) \leq 0, \\ 1 - M(t,y)/\rho, & \text{if } 0 < M(t,y) < \rho, \\ 0, & \text{if } t \geq \rho,\end{cases} \tag{5}$$

where $M(t,y) = t_y - \max_{y' \neq y} t_{y'}$. The margin loss (5) is an upper bound on the zero-one loss and is $\|\cdot\|_\infty$-Lipschitz with constant $\frac{2}{\rho}$ in the first argument for all $y \in \mathcal{Y}$ (Bartlett et al., 2017).

ℓ_∞-additive perturbation We first consider the ℓ_∞ attack (Goodfellow et al., 2014; Kannan et al., 2018), in which the attacker utilizes an additive noise application function $A(x,\delta) = x + \delta$, where $x \in \mathcal{X}$, and the noise set is the ℓ_∞-ball

$$\mathcal{B} = \{\delta : \|\delta\|_\infty \leq \beta\} \subset \mathbb{R}^d.$$

To apply Lemma 4.4, we first show the Lipschitzness of the function $\delta \mapsto \ell_\rho(W(x+\delta), y)$.

Lemma 5.1. *Consider the function $g_W(z,\delta) = \ell(W(x+\delta), y)$ and assume $\|W\|_{1,\infty} \leq \Lambda_1$. Then, for any z, the function $\delta \mapsto g_W(z,\delta)$ is $\|\cdot\|_\infty$-Lipschitz with constant $\frac{2\Lambda_1}{\rho}$.*

Based on the Lipschitzness of the adversarial loss function in the above lemma, we can bound the covering number of the adversarial class in the lemma below.

Lemma 5.2. *Let $\mathcal{F}$ be the multi-class linear hypothesis class defined in (4) and ℓ_ρ the multi-class margin loss (5). Further, let $\mathcal{G}_{\text{adv}}$ and $\tilde{\mathcal{G}}_{\text{adv}}$ be defined as in (1) and (2). Then, for $\epsilon > 0$ and $\tilde{S}$ defined in (3), we have*

$$\log \mathcal{N}_\infty(\epsilon/2, \tilde{\mathcal{G}}_{\text{adv}}, \tilde{S}) \leq \frac{C\Lambda^2(\Psi + \sqrt{d}\beta)^2 L_{\log}}{\epsilon^2\rho^2},$$

where C is an absolute constant, $\Psi' = \Psi + \sqrt{d}\beta$ and

$$L_{\log} := \log\left(2\left\lceil \frac{16\Lambda(\Psi')}{\epsilon\rho} + 2\right\rceil nK\left(\frac{12\beta\Lambda_1}{\rho\epsilon}\right)^d + 1\right).$$

If we plug the above lemma back into Theorem 4.8, we get the following corollary.

Corollary 5.3. *With the notation above, with probability at least $1-\delta$ over the draw of the training data, for all $w \in \mathbb{W}$, we have*

$$R_{\text{adv}}(w) \leq \hat{R}_{\text{adv}}(w) + \frac{8}{n} + 3\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{C\Lambda(\Psi + \sqrt{d}\beta)\tilde{L}_{\log}}{\sqrt{n}\rho}, \quad (6)$$

where C is an absolute constant, $\Psi' = \Psi + \sqrt{d}\beta$, and

$$\tilde{L}_{\log} = \log^{\frac{1}{2}}\left(4\left\lceil \frac{8\Lambda(\Psi')n}{\rho} + 1\right\rceil nK\left(\frac{12\beta\Lambda_1 n}{\rho}\right)^d + 1\right)\log(n).$$

Remark 5.4. We note that the dependence $\sqrt{d}$ on d in the term $\Psi + \sqrt{d}\beta$ is due to the mismatch between the norm on the input x and the norm in the ball $\mathcal{B}$. Indeed, we have used the inequality $\|\delta\|_2 \leq \sqrt{d}\|\delta\|_\infty$. This dependence on d vanishes if the bound on the norms of x and δ matches (e.g., if we consider the attack where $\mathcal{B} := \{\delta : \|\delta\|_2 \leq \beta\}$).

Remark 5.5. The term $\tilde{L}_{\log}$ incurs also a square root dependence on the dimension d. Such a dependence is attributed to the complexity of the adversarial noise set $\mathcal{B}$. For example, if $\mathcal{B}$ is contained in a low dimensional manifold $d' < d$, the dependence is reduced to $O(\sqrt{d'})$. This motivates projecting the input onto a low-dimensional manifold to reduce the effective dimensionality of the adversarial noise.

Remark 5.6. We now compare the bound (6) to the bounds in Yin et al. (2019); Xiao et al. (2021), Khim & Loh (2018), and Awasthi et al. (2020). The dependence of the bound (6) on the number of classes is of the order $O(\log(K))$, while the bounds in Yin et al. (2019); Xiao et al. (2021) and Khim & Loh (2018) are of the order $O(K\sqrt{K})$ and $O(K)$, respectively. Therefore, the bound (6) is favourable in the classification setting with a large K.

The dependence of the bound (6) on the input dimension is of the order $O(d)$. On the other hand, the dependence on d in Yin et al. (2019), Awasthi et al. (2020), and Khim & Loh (2018) is of the order $O(\sqrt{d})$, and thus our bound incurs an extra $\sqrt{d}$ term. Similar to our bound the cost of $\sqrt{d}$ in their bounds is due to the mismatch of the norm constraint on x and δ. Unlike the bound (6), their bounds do not incur the $\sqrt{d}$ resulting from the complexity of the adversarial noise set. This is because for linear models with ℓ_p-additive perturbation the function class transformation in their analysis is tight, and thus effectively the set $\mathcal{B}$ is a singleton with dimension 0. Hence, the bounds (Yin et al., 2019; Awasthi et al., 2020; Khim & Loh, 2018) are favourable for multi-class linear models with ℓ_p-additive perturbation when $K < \sqrt{d}$.

Adversarial spatial transformation We now consider the adversarial spatial transformation in Engstrom et al. (2019). It is based on the spatial transformer network (Jaderberg et al., 2015), which introduced a parameterized spatial transformation that is Lipschitz in the transformation parameters. For simplicity of presentation, we consider square images ($x \in \mathbb{R}^{\sqrt{d}\times\sqrt{d}}$ with $\sqrt{d}$ being an integer) and linear spatial transformations (including rotating and shearing). Therefore, the noise parameter is a matrix $\delta \in \mathbb{R}^{2\times 2}$. The adversarial transformation function A is defined in the following steps: first, the indices of the image are transformed by

$$\begin{pmatrix} u^s \\ v^s \end{pmatrix} = \begin{pmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{pmatrix} \begin{pmatrix} u^t \\ v^t \end{pmatrix}, \quad (7)$$

where $u^t, v^t \in [\sqrt{d}]$ are the target indices of the pixel at the source indices u^s, v^s. Let $U^s \in [\sqrt{d}]^d$ be the aggregation of all the u^s indices in one vector. Define V^s, U^t, V^t similarly. Denote by $(U^s, V^s) = \mathcal{S}(U^t, V^t, \delta)$ the transformation (7).

The output image $\tilde{x}$ is then obtained by setting the value at the index (U_i^t, V_i^t) for $i \in [d]$ according to

$$\sum_{k=1}^{\sqrt{d}} \sum_{l=1}^{\sqrt{d}} x_{kl} \max(0, 1-|V_i^s - k|) \max(0, 1-|U_i^s - l|). \quad (8)$$

Let $\tilde{x} = \mathcal{T}(x, (U^s, V^s))$ denote the transformation (8). The noise application function for the attack A is then defined as

$$A(x, \delta) = \mathcal{T}(x, \mathcal{S}(U^t, V^t, \delta)). \quad (9)$$

The following lemma establishes the Lipschitzness of the functions A and $\delta \mapsto \ell_\rho(WA(x, \delta), y)$.

Lemma 5.7. *Let $A(x, \delta)$ be the noise application function defined above. For all $\|x\|_1 \leq \Psi_1$, the function $\delta \mapsto A(x, \delta)$ is $(\|\cdot\|_\infty, \|\cdot\|_\infty)$-Lipschitz with constant $4\Psi_1\sqrt{d}$. Further, the function $\delta \mapsto \ell_\rho(WA(x, \delta), y)$ is $\|\cdot\|_\infty$-Lipschitz with constant $\frac{8\Psi_1\Lambda_1\sqrt{d}}{\rho}$.*

We can now derive an upper bound on the ℓ_∞-covering number of the adversarial loss. The result is summarized in the following lemma.

Lemma 5.8. *Let $\mathcal{F}$ be the multi-class linear hypothesis class (4) and ℓ_ρ the multi-class margin loss (5). Let $\mathcal{G}_{\text{adv}}$ and $\tilde{\mathcal{G}}_{\text{adv}}$ be defined as (1) and (2), respectively. Then, for $\epsilon > 0$, $\|\delta\|_\infty \leq \beta$ and $\tilde{S}$ defined in (3), we have*

$$\log \mathcal{N}_\infty(\epsilon/2, \tilde{\mathcal{G}}_{\text{adv}}, \tilde{S}) \leq \frac{C\Lambda^2\Psi^2 L_{\log}}{\epsilon^2\rho^2},$$

where C is an absolute constant,

$$L_{\log} := \log\left(2\left\lceil\frac{16\Lambda\Psi}{\epsilon\rho} + 2\right\rceil nK\left(\frac{a}{\rho\epsilon}\right)^4 + 1\right),$$

$a = C_1\beta\Lambda_1\Psi_1\sqrt{d}$, and C_1 is a constant.

Applying Lemma 5.8 and Theorem 4.8, we get the following immediate corollary.

Corollary 5.9. *With the notation above, with probability at least $1-\delta$ over the draw of the training data, we have, for all $W \in \mathbb{W}$,*

$$R_{\text{adv}}(W) \leq \hat{R}_{\text{adv}}(W) + 3\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{8}{n} + \frac{C\Lambda\Psi\tilde{L}_{\log}}{\sqrt{n}\rho},$$

where C is an absolute constant and

$$\tilde{L}_{\log} := \log^{\frac{1}{2}}\left(4\left\lceil\frac{8\Lambda\Psi n}{\rho} + 1\right\rceil nK\left(\frac{an}{\rho}\right)^4 + 1\right)\log(n).$$

Remark 5.10. Note that unlike the bound (6), there is no direct dependence on the input dimension d outside of the log terms. This is because spatial transformations do not alter the image norm. Furthermore, the complexity of the noise set $\mathcal{B}$ is drastically reduced ($\mathcal{B}$ is four-dimensional) in comparison to the additive noise attack. Thus the important factor is the complexity of the adversarial noise set and not the input dimension.

Remark 5.11. To the best of our knowledge, this is the first adversarial generalization bound valid for an attack other than the ℓ_p-additive attack. It is unclear how to adapt the existing approaches in the literature to general attacks.

5.2. Neural Network

We now turn our attention to feed-forward neural networks under different attacks. Recall that a feed-forward network is defined as the composition of a set of L-layers. Each layer $l \in [L]$ consists of a linear transformation parameterized by the matrix $W^l \in \mathbb{R}^{m_l \times m_{l-1}}$, where m_l is the width of layer l, followed by an element-wise non-linear Lipschitz activation function $\sigma(\cdot)$. We have $m_0 = d$ (the input dimension) and $m_L = K$ (the number of classes). Therefore the network function $N_{\mathcal{W}}(x)$ is evaluated as

$$N_{\mathcal{W}}(x) = W^L\sigma(W^{L-1}\sigma(\cdots\sigma(W^1x))\cdots)), \quad (10)$$

where $\mathcal{W} = (W^1, \ldots, W^L)$. We consider norm-bounded networks with the following hypothesis class

$$\mathcal{F} := \{x \mapsto N_{\mathcal{W}}(x) : \mathcal{W} \in \mathbb{W}\}. \quad (11)$$

The quality of classification is measured by the margin loss function (5). In the following, we summarize the assumptions used throughout this section.

Assumption 5.12. Let $\mathcal{W} \in \mathbb{W}$ be the weight of the network (11). Suppose that $\mathbb{W}$ is such that, for all $\mathcal{W} \in \mathbb{W}$, it is $\|W^l\|_2 \leq a_l$ and $\|W^l\|_\sigma \leq s_l$ for all $l \in [L-1]$. Further, suppose that, for all $\mathcal{W} \in \mathbb{W}$, it is $\|W^L\|_2 \in a_L$, $\|W^L\|_{2,\infty} \leq s_L$ and $\|W^1\|_{1,\infty} \leq s_1'$.

ℓ_∞-additive perturbation We now consider the ℓ_∞-additive perturbation applied to multi-layer neural networks. As with the linear case, we first establish the $\|\cdot\|_\infty$-Lipschitzness of the function $\delta \mapsto \ell_\rho(N_{\mathcal{W}}(x+\delta), y)$. The following lemma summarizes the result.

Lemma 5.13. *Let $N_{\mathcal{W}}$ be the neural network function defined in (10). Further let ℓ_ρ be the loss function (5). With Assumption 5.12, the function $g(\delta) = \ell_\rho(N_{\mathcal{W}}(x+\delta), y)$ is $\|\cdot\|_\infty$-Lipschitz in δ with constant $\frac{2}{\rho}(\prod_{l=2}^L s_l)s_1'\sqrt{m_1}$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $\mathcal{W} \in \mathbb{W}$.*

The following lemma introduces an upper bound of the ℓ_∞-covering number of the neural network adversarial class.

Lemma 5.14. *Let $\mathcal{F}$ be the multi-class neural network hypothesis class (11) and ℓ_ρ be the margin loss (5). Let*

$\mathcal{B} := \{\delta : \|\delta\|_\infty \leq \beta\}$. *Further let* $\mathcal{G}_{\text{adv}}$ *and* $\tilde{\mathcal{G}}_{\text{adv}}$ *be defined as* (1) *and* (2)*, respectively. Assume Assumption 5.12 holds and* $\|x\|_2 \leq \Psi$. *Then, for* $\tilde{S}$ *defined in* (3) *and* $\epsilon > 0$*, we have*

$$\log(\mathcal{N}_\infty(\epsilon/2, \tilde{\mathcal{G}}_{\text{adv}}, \tilde{S})) \leq \frac{CL^2\Psi'^2}{\rho^2\epsilon^2}\prod_{l=1}^{L} s_l^2 \left(\sum_{l=1}^{L}\frac{a_l^2}{s_l^2}\right) L_{\log},$$

where

$$L_{\log} := \log\left((C_1\Psi'\Gamma/(\epsilon\rho) + C_2\bar{m})\, n \left(\frac{6\beta\lambda}{\epsilon\rho}\right)^d + 1\right),$$

$\Psi' = (\Psi + \sqrt{d}\beta)$, $\Gamma = \max_{l\in[L]}(\prod_{i=1}^{L} s_i)a_l m_l/s_l$, $\lambda = \frac{2}{\rho}(\prod_{l=2}^{L} s_l)s_1'\sqrt{m_1}$, $\bar{m} = \max_{l\in[L]} m_l$, *and* C, C_1, C_2 *are universal constants.*

The following corollary follows directly from the above Lemma and Theorem 4.8.

Corollary 5.15. *With the notation above, with probability at least* $1-\delta$ *over the draw of the training data, for all* $\mathcal{W} \in \mathbb{W}$*, we have*

$$R_{\text{adv}}(\mathcal{W}) \leq \hat{R}_{\text{adv}}(\mathcal{W}) + \frac{8}{n} + 3\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{CL\Psi'}{\rho\sqrt{n}}\prod_{l=1}^{L} s_l \sqrt{\left(\sum_{l=1}^{L}\frac{a_l^2}{s_l^2}\right)\tilde{L}_{\log}},$$

where C is an absolute constant, $\Psi' = \Psi + \sqrt{d}\beta$ *and*

$$\tilde{L}_{\log} = \log^{\frac{1}{2}}\left(\left(\frac{C_1\Psi'\Gamma n}{\rho} + C_2\bar{m}\right) n \left(\frac{6\beta\lambda n}{\rho}\right)^d + 1\right)\log(n).$$

Remark 5.16. Note that similar to the linear case, the bound has two $\sqrt{d}$ dependence. The first is in Ψ', which arises from the mismatch of norms and can be mitigated by controlling the appropriate norm as discussed above. The second $\sqrt{d}$ term in $\tilde{L}_{\log}$ is due to the complexity of the adversarial noise set $\mathcal{B}$. As discussed in the linear case, a projection on a low-dimensional manifold can help reduce such dependency.

Remark 5.17. The generalization bound in Yin et al. (2019) applies only to a one-hidden-layer neural network and is based on the surrogate upper bound introduced in Raghunathan et al. (2018). This is in contrast to our bound, which directly applies to multi-layer networks and the adversarial loss. While the bound in Khim & Loh (2018) applies to multi-layer networks, it is based on a harsh surrogate upper bound, which pushes the maximization through each layer, thus multiplying the bound slack. This can lead to a vacuous bound for Deep Neural Networks. Furthermore, their bound is of the order $O(K\sqrt{d})$ compared to our bound $O(\log(K)\sqrt{d})$ (using compatible norms on x and δ). While the result in Awasthi et al. (2020) applies directly to the adversarial loss, it applies only to one-hidden-layer neural networks and binary classification. Their bound is of the order $O(\sqrt{d\bar{m}})$ while ours is $O(\sqrt{d}\log(\bar{m}))$, where $\bar{m}$ is the width of the hidden layer.

Adversarial spatial transformation We consider the adversarial attack based on the spatial transformation in (9). We begin by establishing the Lipschtizness of $\delta \mapsto \ell_\rho(N_{\mathcal{W}}(A(x,\delta)), y)$.

Lemma 5.18. *Let* $g(\delta) = \ell_\rho(N_{\mathcal{W}}(A(x,\delta)), y)$ *and Assumption 5.12 hold. Then for all* $\mathcal{W} \in \mathbb{W}$, $(x,y) \in \mathcal{X}\times\mathcal{Y}$, g *is* $\|\cdot\|_\infty$*-Lipschitz with constant* $\frac{8}{\rho}(\prod_{l=2}^{L} s_l)s_1'\sqrt{m_1}\Psi_1\sqrt{d}$.

Lemma 5.19. *Let* $\mathcal{F}$ *be the multi-class neural network hypothesis class* (11) *and* ℓ_ρ *be the margin loss* (5)*. For the adversarial spatial attack defined above, let* $\mathcal{G}_{\text{adv}}$ *and* $\tilde{\mathcal{G}}_{\text{adv}}$ *be defined as in* (1) *and* (2)*, respectively. Suppose that Assumption 5.12 holds, and* $\|x\|_1 \leq \Psi_1$, $\|x\|_2 \leq \Psi$, *and* $\|\delta\|_\infty \leq \beta$*, for all* $\delta \in \mathcal{B}$. *Then for* $\tilde{S}$ *defined in* (3) *and* $\epsilon > 0$*, we have*

$$\log(\mathcal{N}_\infty(\tilde{\mathcal{G}}_{\text{adv}}, \tilde{S}, \epsilon/2)) \leq \frac{CL^2\Psi^2}{\rho^2\epsilon^2}\prod_{l=1}^{L} s_l^2 \left(\sum_{l=1}^{L}\frac{a_l^2}{s_l^2}\right) L_{\log},$$

where

$$L_{\log} := \log\left((C_1\Psi\Gamma/(\epsilon\rho) + C_2\bar{m})\, n \left(\frac{C_3\beta\lambda}{\epsilon\rho}\right)^4 + 1\right),$$

C, C_1, C_2, C_3 *are universal constants,* $\bar{m} = \max_{l\in[L]} m_l$, $\Gamma = \max_{l\in[L]}(\prod_{i=1}^{L} s_i)a_l m_l/s_l$, *and* $\lambda = (\prod_{l=2}^{L} s_l)s_1'\sqrt{m_1}(\Psi_1\sqrt{d})$.

We can plug the above complexity bound back into Theorem 4.8 and derive the following generalization error bounds.

Corollary 5.20. *With the notation above, with probability at least* $1-\delta$ *over the draw of the training data, for all* $\mathcal{W} \in \mathbb{W}$*, we have*

$$R_{\text{adv}}(\mathcal{W}) \leq \hat{R}_{\text{adv}}(\mathcal{W}) + \frac{8}{n} + 3\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{CL\Psi}{\rho\sqrt{n}}\prod_{l=1}^{L} s_l \sqrt{\left(\sum_{l=1}^{L}\frac{a_l^2}{s_l^2}\right)\tilde{L}_{\log}},$$

where C is an absolute constant, and

$$\tilde{L}_{\log} = \log^{\frac{1}{2}}\left(\left(\frac{C_1\Psi\Gamma n}{\rho} + C_2\bar{m}\right) n \left(\frac{6\beta\lambda n}{\rho}\right)^4 + 1\right)\log(n).$$

Remark 5.21. The bound has no direct dependence on d outside of log terms as with the linear case. This is due to the low complexity of the adversarial spatial transform without altering the image norm.

6. Optimistic Bounds and Fast Rates

In this section, we aim to derive optimistic bounds for the generalization of adversarial learning in the sense of incorporating the training errors into the generalization bounds. Optimistic bounds have been studied before in the binary-classification settings (Srebro et al., 2010) and multi-classification settings (Reeve & Kaban, 2020), where they have resulted in fast-rate bounds for smooth losses under low-noise conditions. We aim to extend these approaches to the case of adversarial examples. Our results are based on the *local Rademacher complexity* (Bartlett et al., 2005).

Definition 6.1 (Rademacher complexity)**.** The empirical Rademacher complexity of a function class $\mathcal{F}$ of real-valued functions with respect to a set $S = \{z_i\}_{i=1}^n$ is defined as

$$\mathfrak{R}_S(\mathcal{F}) = \mathbb{E}_\sigma\Big[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\Big], \tag{12}$$

where $\{\sigma_i\}$ are random variables with equal probability of being either $+1$ or -1.

Our result applies loss functions ℓ with a certain smoothness condition defined by the following robust-self-bounding property.

Definition 6.2. Let $\ell : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}$ be a loss function. We say that ℓ is *robust-self-bounding-Lipschitz* with respect to a set $\mathcal{B}$ if, for all $y \in \mathcal{Y}$ and all measurable functions $\nu : \mathcal{B} \to \mathbb{R}^K$ and $\mu : \mathcal{B} \to \mathbb{R}^K$, we have

$$|\max_{\delta\in\mathcal{B}} \ell(\nu(\delta), y) - \max_{\delta\in\mathcal{B}} \ell(\mu(\delta), y)| \le \lambda \max_{\delta\in\mathcal{B}} \|\nu(\delta) - \mu(\delta)\|_\infty \times \max\{\max_{\delta\in\mathcal{B}} \ell(\nu(\delta), y), \max_{\delta\in\mathcal{B}} \ell(\mu(\delta), y)\}^\theta.$$

Remark 6.3. The robust self-bounding property is inspired by the *self-bounding* property introduced in Reeve & Kaban (2020) as follows

$$|\ell(\nu, y) - \ell(\mu, y)| \le \lambda \max\{\ell(\nu, y), \ell(\mu, y)\}^\theta \|\nu - \mu\|_\infty.$$

The key difference is the introduction of the adversarial max operator $\max_{\delta\in\mathcal{B}}$ to adapt to adversarial learning.

Remark 6.4. The robust-self-bounding property is maintained by several realistic losses. For instance, consider the smooth margin loss $L_\rho(t, y) =$

$$\begin{cases} 1 & \text{if } M(t,y) \le 0 \\ 2\left(M(t,y)/\rho\right)^3 - 3\left(M(t,y)/\rho\right)^2 + 1 & \text{if } 0 < M(t,y) < \rho \\ 0 & \text{if } M(t,y) \ge \rho, \end{cases}$$

defined in Reeve & Kaban (2020). This function is an upper bound on the zero-one loss, therefore, it can be considered as a surrogate loss to classification scenarios. It is further a robust-self-bounding-Lipschitz with $\theta = 1/2$ and $\lambda = 4\sqrt{6}/\rho$ as shown in appendix E.

We now define the local hypothesis class and the local loss class. For a given hypothesis class

$$\mathcal{F}_{\text{adv}} := \big\{(x, \delta) \mapsto f(A(x, \delta), w) : w \in \mathbb{W}\big\},$$

we define the local hypothesis class $\mathcal{F}_{\text{adv}}|^r \subset \mathcal{F}_{\text{adv}}$ as the set of functions $f \in \mathcal{F}_{\text{adv}}$ with empirical adversarial training errors at most r. That is

$$\mathcal{F}_{\text{adv}}|^r := \Big\{(x, \delta, y) \mapsto f(A(x, \delta), w) : w \in \mathbb{W}, \hat{R}_{\text{adv}}(w) \le r\Big\}.$$

Similarly, define the local loss class

$$\mathcal{G}_{adv}|^r := \Big\{(x, y) \mapsto \max_{\delta\in\mathcal{B}} \ell(f(A(x, \delta), w), y) : w \in \mathbb{W}, \hat{R}_{\text{adv}}(w) \le r\Big\}.$$

We first introduce a structural result on covering numbers.

Lemma 6.5. *Let $\mathcal{G}_{adv}$ be defined as above. Assume that the loss ℓ is robust-self-bounding with parameters $\lambda > 0, \theta \in (0, 1/2]$. Further let $\delta \mapsto \ell(f(A(x, \delta), w), y)$ be $\|\cdot\|$-Lipschitz with constant L. Let $\tilde{\mathcal{F}}_{\text{adv}} := \{(x, \delta, y) \mapsto f(A(x, \delta), w)_y : w \in \mathbb{W}\}$ and $\hat{S} := \{(x_i, \delta, \tilde{y}) : i \in [n], \delta \in C_{\mathcal{B}}(\frac{\epsilon}{\lambda(2r)^\theta 2L}), \tilde{y} \in \mathcal{Y}\}$, where $C_{\mathcal{B}}(\epsilon/2L)$ is an $(\epsilon/2L, \|\cdot\|)$-cover of $\mathcal{B}$. Then, we have*

$$\mathcal{N}_2(\epsilon, \mathcal{G}_{adv}|^r, S) \le \mathcal{N}_\infty\left(\frac{\epsilon}{4\lambda(2r)^\theta}, \tilde{\mathcal{F}}_{\text{adv}}, \hat{S}\right).$$

Lemma 6.5 establishes a bound on the ℓ_2-covering number of the local loss class by the ℓ_∞-covering number of the extended hypothesis class $\tilde{\mathcal{F}}_{\text{adv}}$. This serves as a key step in the proof of the next lemma, which establishes a bound on the local Rademacher complexity by a sub-root function of r, a key step in developing optimistic bounds (Bartlett et al., 2005; Srebro et al., 2010).

Lemma 6.6. *With the notation and assumptions of Lemma 6.5, suppose that, for all $w \in \mathbb{W}$, $\|f(x, w)\|_\infty \le B$ and ℓ is bounded by b. Suppose that $\log \mathcal{N}_\infty\left(\epsilon, \tilde{\mathcal{F}}_{\text{adv}}, \hat{S}\right) \le \frac{R_{b_1}}{\epsilon^2}$, for $\epsilon \in [b_1, b_2]$ and $R_{b_1} \in \mathbb{R}$ that does not depend on n and ϵ. Then,*

$$\mathfrak{R}_S(\mathcal{G}_{adv}|^r) \le \frac{\lambda r^\theta}{n}\left[2^{4+\theta} + 40\sqrt{R_{\frac{1}{\sqrt{n}}}} \log\left(\frac{b^{1-\theta}\sqrt{n}}{2^{2+\theta}\lambda}\right)\right].$$

Lemma 6.6 establishes a bound on the Rademacher complexity of the loss class $\mathcal{G}_{\text{adv}}|^r$ in terms of a bound on the empirical risk r. Note that we obtain a sub-root bound on the local Rademacher complexity for $\theta = 1/2$.

Remark 6.7. The condition $\log \mathcal{N}_\infty \left(\epsilon, \tilde{\mathcal{F}}_{\text{adv}}, \hat{S}\right) \leq \frac{R_{b_1}}{\epsilon^2}$ is satisfied by many of the typical function calsses. For instance, it is satified by the neural network class (see Lemma C.1) with

$$R_{b_1} = CL^2\Psi'^2 \prod_{l=1}^{L} s_l^2 \left(\sum_{l=1}^{L} \frac{a_l^2}{s_l^2}\right) L_{\log},$$

where

$$L_{\log} := \log\left(\left(C_1\Psi'\Gamma/b_1\right) + C_2\bar{m}\right) n \left(\frac{6\beta\lambda}{b_1}\right)^d + 1\Bigg).$$

The next theorem presents the main result of this section, namely an optimistic generalization bounds for adversarial learning with a smooth loss ℓ.

Theorem 6.8. *With the notation and assumption of Lemmas 6.6 and 6.5, with probability at least $1-\delta$ over the draw of the training set S, for all $w \in \mathbb{W}$, we have $R_{\text{adv}}(w)$ is bounded by*

$$R_{\text{adv}}(w) \leq \hat{R}_{\text{adv}}(w) + 106r + \sqrt{\hat{R}_{\text{adv}}(w)\,(8r+L)} + \frac{48b}{n}(\log(1/\delta)\log(\log(n))),$$

where $L = \frac{4b}{n}(\log(1/\delta) + \log(\log(n)))$ and

$$r = \frac{\lambda^2}{n}\left[16\sqrt{2} + 40\sqrt{R_{\frac{1}{\sqrt{n}}}}\log\left(\frac{b^{1-\theta}\sqrt{n}}{2^{2+\theta}\lambda}\right)\right]^2.$$

Remark 6.9. Note that the second term grows as $\tilde{O}(\frac{R_{1/\sqrt{n}}}{n})$. For the majority of function classes, $R_{1/\sqrt{n}}$ is $\tilde{O}(1)$, where $\tilde{O}$ hides log terms, for example for linear models (Lei et al., 2019)), kernel methods (Bartlett & Mendelson, 2003), DNNs (Bartlett et al., 2017), and structured output prediction (Mustafa et al., 2021). The second term would then grow as $\tilde{O}(\frac{1}{n})$. The fourth term grows at the usual $\tilde{O}(\frac{1}{\sqrt{n}})$ rate. However, if $\hat{R}_{\text{adv}}(w) = 0$, the fourth term vanishes, thus leading to a $\tilde{O}(\frac{1}{n})$ generalization bound.

Remark 6.10. In Dan et al. (2020), the excess risk bound of the order $O(\frac{d}{n})$ was shown under an assumption on the adversarial signal-to-noise ratio. However, it applies to linear classes in the idealized case that the true data distribution is Gaussian. On the other hand, our bound applies to any function class where the covering number or the Rademacher complexity can be bounded. Bhattacharjee et al. (2021) presented bounds in expectation of the order $O(1/n)$ for linear and kernel-based models under the distributional assumptions of separable data. In comparison, we develop high-probability bounds applicable to any function class under milder distributional assumptions (zero empirical risk).

7. Conclusion

We presented a general generalization analysis of adversarial learning. Our analysis applies to a wide range of attacks and models. To our knowledge, this is the first analysis for non-additive noise. Our approach is modular and easily applicable to a large number of models. We showcased our general results for linear models and neural networks with additive noise or the spatial transformation attack. Our analysis emphasized the importance of the complexity of the adversarial noise set $\mathcal{B}$ rather than the input space dimension. We further extended our analysis to the case of smooth losses, where we derived fast-rates under zero adversarial empirical risk. In future work, we will investigate mitigating the dependence on the dimension of the noise set $\mathcal{B}$. Our hypothesis is that the complexity can be mitigated by carefully considering the interplay between the noise set and the function class (e.g., an additive attack on a model first reducing the input dimension to $d' < d$ should yield bounds with a dependence on d', rather than d).

Acknowledgements

We thank the reviewers for their constructive feedback that helped improve the paper. MK and WM acknowledge support by the Carl-Zeiss Foundation, the DFG awards KL 2698/2-1 and KL 2698/5-1, and the BMBF awards 01|S18051A, 03|B0770E, and 01|S21010C.

References

Athalye, A., Carlini, N., and Wagner, D. A. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018.

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

Awasthi, P., Frank, N., and Mohri, M. Adversarial learning guarantees for linear hypotheses and neural networks. In *International Conference on Machine Learning*, pp. 431–441. PMLR, 2020.

Awasthi, P., Yu, G., Ferng, C.-S., Tomkins, A., and Juan, D.-C. Adversarial robustness across representation spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7608–7616, 2021.

Barreno, M., Nelson, B., Joseph, A. D., and Tygar, J. D. The security of machine learning. *Machine Learning*, 81 (2):121–148, 2010.

Bartlett, P., Foster, D., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30: 6241–6250, 2017.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. 3(null): 463–482, mar 2003. ISSN 1532-4435.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local rademacher complexities. *The Annals of Statistics*, 33(4): 1497–1537, 2005.

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

Bietti, A., Mialon, G., Chen, D., and Mairal, J. A kernel perspective for regularizing deep neural networks. *arXiv preprint arXiv:1810.00363*, 2018.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 387–402. Springer, 2013.

Boucheron, S., Lugosi, G., and Bousquet, O. Concentration inequalities. In *Summer school on machine learning*, pp. 208–240. Springer, 2003.

Brendel, W., Rauber, J., and Bethge, M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

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

Carlini, N., Katz, G., Barrett, C., and Dill, D. L. Ground-truth adversarial examples. *CoRR*, abs/1709.10207, 2017. URL http://arxiv.org/abs/1709.10207.

Cissé, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. In *ICML*, 2017.

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

Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., and Madry, A. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*, pp. 1802–1811. PMLR, 2019.

Farnia, F., Zhang, J., and Tse, D. Generalizable adversarial training via spectral normalization. In *International Conference on Learning Representations*, 2018.

Finlay, C. and Oberman, A. M. Scaleable input gradient regularization for adversarial robustness. *arXiv preprint arXiv:1905.11468*, 2019.

Gao, Q. and Wang, X. Theoretical investigation of generalization bounds for adversarial learning of deep neural networks. *Journal of Statistical Theory and Practice*, 15 (2):1–28, 2021.

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

Hein, M. and Andriushchenko, M. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *NIPS*, 2017.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

Jaderberg, M., Simonyan, K., Zisserman, A., et al. Spatial transformer networks. *Advances in Neural Information Processing Systems*, 28:2017–2025, 2015.

Kannan, H., Kurakin, A., and Goodfellow, I. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

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

Ledent, A., Alves, R., Lei, Y., and Kloft, M. Fine-grained generalization analysis of inductive matrix completion. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 25540–25552. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper/2021/file/d6428eecbe0f7dff83fc607c5044b2b9-Paper.pdf.

Ledent, A., Mustafa, W., Lei, Y., and Kloft, M. Norm-based generalisation bounds for deep multi-class convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 9 in 35, pp. 8279–8287, 2021b.

Lei, Y., Dogan, Ü., Zhou, D.-X., and Kloft, M. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65(5): 2995–3021, 2019.

Long, P. M. and Sedghi, H. Size-free generalization bounds for convolutional neural networks. In *ICLR*, 2020.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *CoRR*, abs/1706.06083, 2017.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Mustafa, W., Vandermeulen, R. A., and Kloft, M. Input hessian regularization of neural networks. In *Workshop on "Beyond first-order methods in ML systems" at the 37th International Conference on Machine Learning*, 2020.

Mustafa, W., Lei, Y., Ledent, A., and Kloft, M. Fine-grained generalization analysis of structured output prediction. In *IJCAI 2021*, 2021.

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

Prabhu, Y. and Varma, M. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 263–272, New York, NY, USA, 2014.

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. *CoRR*, abs/1801.09344, 2018.

Reeve, H. W. and Kaban, A. Optimistic bounds for multi-output prediction. In *37th International Conference on Machine Learning*, 2020.

Ross, A. S. and Doshi-Velez, F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *NeurIPS*, 2018.

Srebro, N., Sridharan, K., and Tewari, A. Smoothness, low noise and fast rates. *Advances in Neural Information Processing Systems*, 23, 2010.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

Wong, E. and Kolter, J. Z. Learning perturbation sets for robust machine learning. In *International Conference on Learning Representations*, 2020.

Wu, L., Ledent, A., Lei, Y., and Kloft, M. Fine-grained generalization analysis of vector-valued learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10338–10346, 2021.

Xiao, J., Fan, Y., Sun, R., and Luo, Z.-Q. Adversarial Rademacher complexity of deep neural networks. 2021.

Xing, Y., Song, Q., and Cheng, G. On the algorithmic stability of adversarial training. *Advances in Neural Information Processing Systems*, 34, 2021.

Yin, D., Kannan, R., and Bartlett, P. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pp. 7085–7094. PMLR, 2019.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.

Zhang, T. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.

A. Proofs of the Main Result in Section 4

In this section, we present the proofs of our main result (Theorem 4.8). As discussed in the main text, our main tool is the ℓ_∞-covering numbers (see Definition 4.1). We note that the main hardness in deriving bounds over the adversarial loss is the maximization over the adversarial noise set. Our main strategy is to utilize the properties of the ℓ_∞-covering numbers to control the complexity of the adversarial loss class with a high function dimensional class, thus alleviating the maximization operator.

Our first step is to bound the covering number of the adversarial class with an extended function class where each element in the adversarial loss class is replaced with a full-function that takes the adversarial noise as its argument as summarized in Lemma 4.2. We now present the proof of Lemma 4.2.

Proof of Lemma 4.2. Our idea is to construct a cover for the adversarial class $\mathcal{G}_{\text{adv}}$ on the training set S from the elements of cover for the class $\mathcal{G}$ on the same training set. To avoid clutter let $g(z,\delta,w) := \ell(f_w(A(x,\delta),y)$. Recall from Definition 4.1 that $\mathcal{N}_\infty(\epsilon,\mathcal{G}_{adv},S)$ is the cardinality of the smallest cover for the set

$$\mathcal{G}_{adv}|_S := \left\{\left(\max_{\delta\in\mathcal{B}} g(z_1,\delta,w),\ldots,\max_{\delta\in\mathcal{B}} g(z_n,\delta,w)\right) : w\in\mathbb{W}\right\} \subset \mathbb{R}^n.$$

The first step of our proof is the observation that it is possible to construct an ℓ_∞-cover for $\mathcal{G}_{adv}|_S$ by utilizing an ℓ_∞-cover constructed for the set

$$\mathcal{G}|_S = \{(g(z_1,\cdot,w),\ldots,g(z_n,\cdot,w))) : w\in\mathbb{W}\} \subset (\mathbb{R}^{\mathcal{B}})^n.$$

Note that each element of g is a vector of functions $g(z,\cdot,w)$ as compared to only the scalar $\max_{\delta\in\mathcal{B}} g(z,\delta,w)$ in $\mathcal{G}_{adv}$. Our aim now is that given a cover for $\mathcal{G}$; we construct a cover for $\mathcal{G}_{adv}$. To that extent, let

$$\mathcal{C}_{\mathcal{G}} = \left\{\left(c_i^1(\cdot),\ldots,c_i^n(\cdot)\right) : i\in[m]\right\} \subset (\mathbb{R}^{\mathcal{B}})^n$$

be a (ϵ,ℓ_∞)-cover of $\mathcal{G}$. We now claim that the set

$$\mathcal{C}_{\mathcal{G}_{adv}} = \left\{\left(\tilde{c}_i^1 := \max_{\delta\in\mathcal{B}} c_i^1(\delta),\ldots,\tilde{c}_i^n := \max_{\delta\in\mathcal{B}} c_i^n(\delta)\right) : i\in[m]\right\}$$

covers the set $\mathcal{G}_{adv}$. Indeed, given $w\in\mathbb{W}$, there exist by definition $j(w)$ such that:

$$\max_{i\in[n]}\max_{\delta\in\mathcal{B}} |g(z_i,\delta,w) - c^i_{j(w)}(\delta)| \le \epsilon.$$

Therefore, we have

$$\begin{aligned}\max_{i\in[n]} |\max_{\delta\in\mathcal{B}} g(z_i,\delta,w) - \tilde{c}^i_{j(w)}| &= \max_{i\in[n]} |\max_{\delta\in\mathcal{B}} g(z_i,\delta,w) - \max_{\delta\in\mathcal{B}} c^i_{j(w)}(\delta)| \\ &\le \max_{i\in[n]}\max_{\delta\in\mathcal{B}} |g(z_i,\delta,w) - c^i_{j(w)}(\delta)| \\ &\le \epsilon.\end{aligned}$$

The first equality follows from the construction of $\mathcal{C}_{\mathcal{G}_{\text{adv}}}$. The first inequality follows from the following inequality: for real-valued functions f, and g, we have $|\max_x f(x) - \max_x g(x)| \le \max_x |f(x) - g(x)|$. It then follows that $\mathcal{C}_{\mathcal{G}_{adv}}$ is an (ϵ,ℓ_∞)-cover to $\mathcal{G}_{adv}$. Since the cardinality of $\mathcal{C}_{\mathcal{G}}$ and $\mathcal{C}_{\mathcal{G}_{adv}}$ are equal, we have

$$\mathcal{N}_\infty(\epsilon,\mathcal{G}_{\text{adv}},S) \le \mathcal{N}_\infty(\epsilon,\mathcal{G},S). \tag{13}$$

The proof is completed. □

Next, we present the proof of Lemma 4.4.

Proof of Lemma 4.4. Our goal is to control the ℓ_∞-covering number of the infinite-dimensional class $\mathcal{G}$ with a finite-dimensional counterpart. The core idea is to approximate the functions $g \in \mathcal{G}$ with a discrete version of it. We, therefore, require the functions in $\mathcal{G}$ to be Lipschitz with respect to some norm $\|\cdot\|$ and a cover of $\mathcal{B}$ with respect to the same norm. To that end, let $\mathcal{C}_\mathcal{B}(\epsilon/2L, \|\cdot\|)$ be an $\epsilon/2L$-cover w.r.t. a general norm $\|\cdot\|$ for the set $\mathcal{B}$. Let $M_\mathcal{B}$ be the size of $\mathcal{C}_\mathcal{B}(\epsilon/2L, \|\cdot\|)$. That is,

$$\mathcal{C}_B := \mathcal{C}_\mathcal{B}(\epsilon/2L, \|\cdot\|) = \left\{\tilde{\delta}_i, i \in [M_\mathcal{B}]\right\} \subset \mathcal{B}.$$

Now recall that $\mathcal{N}_\infty(\epsilon, \tilde{\mathcal{G}}_{\text{adv}}, \tilde{S})$ is the smallest cardinality of the set covering $\mathcal{G}_{\text{disc}}|S$ defined according to the set $\mathcal{C}_\mathcal{B}(\epsilon/2L, \|\cdot\|)$ as follows,

$$\mathcal{G}_{\text{disc}}|_S = \left\{ \begin{pmatrix} g(z_1, \tilde{\delta}_1, w) & g(z_1, \tilde{\delta}_2, w) & \cdots & g(z_1, \tilde{\delta}_{M_\mathcal{B}}, w) \\ \vdots & \vdots & \ddots & \vdots \\ g(z_n, \tilde{\delta}_1, w) & g(z_n, \tilde{\delta}_2, w) & \cdots & g(z_n, \tilde{\delta}_{M_\mathcal{B}}, w) \end{pmatrix} : w \in \mathbb{W} \right\} \subset \mathbb{R}^{n \times M_\mathcal{B}}$$

Our goal now is to construct an (ϵ, ℓ_∞)-cover of $\mathcal{G}|_S$ by utilizing an $(\epsilon/2, \ell_\infty)$-cover of $\mathcal{G}_{\text{disc}}|_S$. To that extent, let the set

$$\mathcal{C}_{\mathcal{G}_{\text{disc}}|_S} = \left\{ \begin{pmatrix} \hat{c}_i^1(\tilde{\delta}_1) & \hat{c}_i^1(\tilde{\delta}_2) & \cdots & \hat{c}_i^1(\tilde{\delta}_{M_\mathcal{B}}) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{c}_i^n(\tilde{\delta}_1) & \hat{c}_i^n(\tilde{\delta}_2) & \cdots & \hat{c}_i^n(\tilde{\delta}_{M_\mathcal{B}}) \end{pmatrix} : i \in [M_{\mathcal{G}_{\text{disc}}|_S}] \right\} \subset \mathbb{R}^{n \times M_\mathcal{B}}.$$

be an $\epsilon/2$-cover for $\mathcal{G}_{\text{disc}}|_S$ of size $M_{\mathcal{B}_{\text{disc}}|_S}$. We now construct a cover of $\mathcal{G}|_S$. The key idea here is to construct functions $c_i(\cdot)$, $i \in [M_{\mathcal{G}_{\text{disc}}|_S}]$ to be piece-wise constant around each $\tilde{\delta}_j$, for $j \in [M_\mathcal{B}]$. Consider the set

$$\mathcal{C}_\mathcal{G} = \left\{ \begin{pmatrix} c_i^1(\delta) := \hat{c}_i^1(\arg\min_{\delta' \in C_\mathcal{B}} \|\delta - \delta'\|) \\ \vdots \\ c_i^n(\delta) := \hat{c}_i^n(\arg\min_{\delta' \in C_\mathcal{B}} \|\delta - \delta'\|) \end{pmatrix} : i \in [M_{\mathcal{G}_{disc}}] \right\} \in (\mathbb{R}^\mathcal{B})^n.$$

We claim that it is an ϵ-cover for $\mathcal{G}$. Indeed, for any $w \in \mathbb{W}$, by construction of $C_{\mathcal{G}_{\text{disc}}|_S}$, there exist $j(w)$, such that

$$\max_i \max_{\tilde{\delta} \in \mathcal{C}_\mathcal{B}(\epsilon/2L)} |g(z_i, \tilde{\delta}, w) - \hat{c}^i_{j(w)}(\tilde{\delta})| \leq \epsilon/2. \tag{14}$$

Therefore,

$$\begin{aligned} \max_i \max_{\delta \in \mathcal{B}} |g(z_i, \delta, w) - c^i_{j(w)}(\delta)| &= \max_i \max_{\delta \in \mathcal{B}} |g(z_i, \delta, w) - g(z_i, \delta^*(\delta), w) + g(z_i, \delta^*(\delta), w) - c^i_{j(w)}(\delta)| \\ &\leq \max_i \max_{\delta \in \mathcal{B}} \Big(|g(z_i, \delta, w) - g(z_i, \delta^*(\delta), w)| + |g(z_i, \delta^*(\delta), w) - c^i_{j(w)}(\delta)| \Big), \end{aligned}$$

where $\delta^*(\delta) := \arg\min_{\delta' \in \mathcal{C}_\mathcal{B}(\epsilon/2L)} \|\delta - \delta^*\|$. The inequality follows from triangle inequality. Since by construction of $\mathcal{C}_\mathcal{G}$ we have $c^i_{j(w)}(\delta) = \hat{c}^i_{j(w)}(\delta^*(\delta))$[2], we have

$$\begin{aligned} \max_i \max_{\delta \in \mathcal{B}} \Big(|g(z_i, \delta, w) - g(z_i, \delta^*(\delta), w)| + |g(z_i, \delta^*(\delta), w) - c^i_{j(w)}(\delta)| \Big) \\ &= \max_i \max_{\delta \in \mathcal{B}} \Big(|g(z_i, \delta, w) - g(z_i, \delta^*(\delta), w)| + |g(z_i, \delta^*(\delta), w) - \hat{c}^i_{j(w)}(\delta^*(\delta))| \Big) \\ &\leq \max_i \max_{\delta \in \mathcal{B}} |g(z_i, \delta, w) - g(z_i, \delta^*(\delta), w)| + \max_i \max_{\tilde{\delta} \in \mathcal{C}_\mathcal{B}} |g(z_i, \tilde{\delta}, w) - \hat{c}^i_{j(w)}(\tilde{\delta})| \\ &\leq \max_i \max_{\delta \in \mathcal{B}} L\|\delta - \delta^*(\delta)\| + \max_i \max_{\tilde{\delta} \in \mathcal{C}_\mathcal{B}} |g(z_i, \tilde{\delta}, w) - \hat{c}^i_{j(w)}(\tilde{\delta})| \\ &\leq L\epsilon/2L + \epsilon/2 = \epsilon \end{aligned}$$

[2]Ties are resolved in arbitrary but fixed manner

The second inequality follows from the Lipschtizness of g with respect to δ and the $\|\cdot\|$-norm. The third inequality is due to the construction of $\delta^*(\cdot)$ and equation (14). Since $\mathcal{C}_{\mathcal{G}_{disc|S}}$ and $\mathcal{C}_{\mathcal{G}}$ have equal size, we conclude that

$$\mathcal{N}_\infty(\epsilon, \mathcal{G}, S) \leq \mathcal{N}_\infty(\epsilon, \mathcal{G}_{\text{disc}}, \tilde{S}). \tag{15}$$

By combining inequalities (13) and (15), the result follows. □

Now we proceed to prove Theorem 4.8. We first present the Rademacher theorem, which controls the generalization of learning algorithms by the Rademacher complexity.

Theorem A.1 (Mohri et al. 2018)**.** *Let $S = \{z_i\}_{i=1}^m$ be i.i.d. random sample from a distribution $\mathcal{D}$ defined over $\mathcal{Z}$. Further let $\mathcal{F} \subset [0,1]^{\mathcal{Z}}$ be a loss class parameterized by the set $\mathbb{W}$. Then for all $\delta \in (0,1)$, we have with probability at least $1-\delta$ over the draw of the sample S, for all $w \in \mathbb{W}$ that*

$$R(w) \leq \hat{R}(w) + 2\mathfrak{R}_S(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}.$$

The theorem states that the generalization of the function class $\mathcal{F}$ can be controlled, with high probability, by the empirical risk and the Rademacher complexity.

Our approach relies, however, on another complexity measure, namely ℓ_∞-covering numbers. The following classical result of Dudley's entropy integral (Boucheron et al., 2003; Bartlett et al., 2017; Ledent et al., 2021a; Srebro et al., 2010) gives a relationship between the Rademacher complexity and ℓ_∞-covering number. We apply the version by Srebro et al. (2010).

Theorem A.2 (Srebro et al. 2010)**.** *Let $\mathcal{F}$ be a class of functions mapping from a space $\mathcal{Z}$ and taking values in $[0,b]$, and assume that $0 \in \mathcal{F}$. Let S be a finite sample of size m and $\hat{\mathbb{E}}[f(z)^2] = \frac{1}{m}\sum_{i=1}^m f(z_i)^2$ We have the following relationship between the empirical Rademacher complexity $\mathfrak{R}_S(\mathcal{F})$ and the covering number $\mathcal{N}_2(\epsilon, \mathcal{F}, S)$.*

$$\mathfrak{R}(\mathcal{F}) \leq \inf_{\alpha>0}\left(4\alpha + \frac{10}{\sqrt{n}}\int_\alpha^{\sup_{f\in\mathcal{F}}\sqrt{\hat{\mathbb{E}}[f(z)^2]}} \sqrt{\log \mathcal{N}_2(\epsilon, \mathcal{F}, S)}d\epsilon\right).$$

We are now ready to present the proof of Theorem 4.8.

Proof of Theorem 4.8. The proof is a direct application of Theorems A.2 and A.1 combined with Lemma 4.2 and Lemma 4.4. For $\delta \in (0,1)$, we have with probability at least $1-\delta$, for all $w \in \mathbb{W}$

$$\begin{aligned}
R_{\text{adv}}(w) &\leq \hat{R}_{\text{adv}}(w) + 2\mathfrak{R}_S(\mathcal{G}_{\text{adv}}) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \\
&\leq \hat{R}_{\text{adv}}(w) + \inf_{\alpha>0}\left(8\alpha + \frac{24}{\sqrt{n}}\int_\alpha^1 \sqrt{\log \mathcal{N}_\infty(\epsilon, \mathcal{G}_{\text{adv}}, S)}d\epsilon\right) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \\
&\leq \hat{R}_{\text{adv}}(w) + \inf_{\alpha>0}\left(8\alpha + \frac{24}{\sqrt{n}}\int_\alpha^1 \sqrt{\log \mathcal{N}_\infty(\epsilon/2, \tilde{\mathcal{G}}_{\text{adv}}, \tilde{S})}d\epsilon\right) + 3\sqrt{\frac{\log(2/\delta)}{2n}}.
\end{aligned}$$

The first inequality is due to Theorem A.1 while the second is derived from Theorem A.2. The third inequality follows from Lemmas 4.2 and 4.4. □

B. Proofs of Results on Linear Models in Section 5.1

In this section, we present the omitted proof of section 5.1. The first step of our approach is to show that the loss function is ℓ_∞-Lipschtiz with respect to the noise parameter δ. We then derive a bound on the size of the set $C_{\mathcal{B}}(\epsilon/2L)$. We finally bound the ℓ_∞-covering number of the extended class $\tilde{\mathcal{G}}_{\text{adv}}$ on the extended data set $\tilde{S}$.

Throughout the paper, we will require upper bounds on the covering number of bounded balls in $\mathbb{R}^d$. We begin by reviewing the following result deriving an upper bound on the size of the set $C_{\mathcal{B}}(\epsilon)$ defined for the general norm $\|\cdot\|$.

Lemma B.1 (Long & Sedghi 2020). *Let d be a positive integer, $\|\cdot\|$ be a norm, ρ be the metric induced by it, and $\kappa, \epsilon > 0$. A ball of radius κ in $\mathbb{R}^d$ w.r.t. ρ can be covered by $(\frac{3\kappa}{\epsilon})^d$ balls of radius ϵ.*

Our approach relies on ℓ_∞-covering numbers for the loss classes. We now review the upper bounds on the ℓ_∞-covering numbers of linear models. The bound used in our work is a combination of the approach in Lei et al. (2019) and the covering number bound in Zhang (2002).

Lemma B.2 (Lei et al. 2019). *Let $\mathcal{F} = \{x \mapsto (w_1^\top x, w_2^\top x, \ldots, x_K^\top x) : \tilde{w} = (w_1, \ldots, w_K) \in \mathbb{R}^{Kd}, \|\tilde{w}\|_2 \leq \Lambda\}$ be the linear model hypothesis class, $S = \{(x_i, y_i)\}_{i=1}^n$ be a given data set. Consider an $\|\cdot\|_\infty$-Lipschitz loss ℓ with constant L, for all $y \in [K]$. Let $\tilde{S}$ be an extended data set defined as $\tilde{S} = \{\phi_j(x_i) : i \in [n], j \in [K]\}$ where $\phi_j(\cdot)$ is defined as*

$$\phi_j(x) := \big(\underbrace{0, \ldots, 0}_{j-1}, x, \underbrace{0, \ldots, 0}_{K-j}\big) \in \mathbb{R}^{Kd}.$$

Further let $\tilde{\mathcal{F}} := \{x \mapsto \langle \tilde{w}, x\rangle, x \in \tilde{S}, \tilde{w} \in \mathbb{R}^{Kd}, \|\tilde{w}\|_2 \leq \Lambda\}$. For all $\epsilon > 0$, we have the following inequality

$$\mathcal{N}_\infty(\epsilon, \ell \circ \mathcal{F}, S) \leq \mathcal{N}_\infty(\epsilon/L, \tilde{\mathcal{F}}, \tilde{S}).$$

The above lemma allows for controlling the ℓ_∞-covering number of multi-class loss classes by a real-valued function class $\tilde{\mathcal{F}}$ on an extended dataset $\tilde{S}$. The following lemma gives a covering number bound for real-valued linear function classes.

Lemma B.3 (Zhang 2002). *Let $\mathcal{L}$ be a class of linear functions on a set of size n. That is, $\mathcal{L} = \{\langle w, x\rangle, x, w \in \mathbb{R}^N\}$. If $\|x\|_q \leq b$ and $\|w\|_p \leq a$, where $2 \leq q < \infty$ and $1/p + 1/q = 1$, then $\forall \epsilon > 0$,*

$$\log \mathcal{N}_\infty(\epsilon, \mathcal{L}, n) \leq 36(q-1)\frac{a^2 b^2}{\epsilon^2} \log[2\lceil 4ab/\epsilon + 2\rceil n + 1],$$

where $\mathcal{N}_\infty(\epsilon, \mathcal{L}, n)$ is the worst case covering number of the class $\mathcal{L}$ on a dataset of size n.

The above result controls the covering numbers by norms of the data and weights. A direction application of Lemmas B.3 and B.2 gives the following corollary.

Corollary B.4. *Let $\mathcal{F}$ be the linear multi-class linear hypothesis class, ℓ_ρ be the loss (5). Let $S = \{(x_i, y_i)\}_{i=1}^n$ be a given dataset with $\|x\|_2 \leq \Psi$, for all $x \in \mathcal{X}$, and $\|W\|_{2,2} \leq \Lambda$, then for all $\epsilon > 0$, we have*

$$\log \mathcal{N}_\infty(\epsilon, \ell_\rho \circ \mathcal{F}, S) \leq C\frac{\Psi^2\Lambda^2}{\rho^2\epsilon^2} \log\left(2\lceil 8\Psi\Lambda/\epsilon\rho + 2\rceil nK + 1\right).$$

Proof. We apply Lemmas B.2 and B.3 noting that ℓ_ρ is $\|\cdot\|_\infty$-Lipschitz with constant $\frac{2}{\rho}$-norm, for all $y \in \mathcal{Y}$. □

B.1. ℓ_∞-additive Perturbation Attack

We are now ready to prove the bounds of the ℓ_∞-additive attacks applied to linear models. Our first step is to derive the $\|\cdot\|_\infty$-Lipschitz constant of the function $\delta \mapsto \ell_\rho(W(x+\delta))$. The following is the poof of Lemma 5.1.

Proof of Lemma 5.1. The proof is a direct derivation. Observe the following, for all $(x, y) \in \mathcal{Z}$ and $\|W\|_{1,\infty} \leq \Lambda_1$, and $\delta, \delta' \in \mathcal{B}$, we have

$$\begin{aligned}|\ell_\rho(W(x+\delta), y) - \ell_\rho(W(x+\delta'), y)| &\leq \frac{2}{\rho}\|W\delta - W\delta'\|_\infty \leq \frac{2}{\rho}\max_i |W_{i,\cdot}\delta - W_{i,\cdot}\delta'| \\ &\leq \frac{2}{\rho}\max_i \|W_{i,\cdot}\|_1 \|\delta - \delta'\|_\infty \leq \frac{2\Lambda_1}{\rho}\|\delta - \delta'\|_\infty,\end{aligned}$$

where $W_{i,\cdot}$ denotes the i-th row of W. The first inequality is derived from that ℓ_ρ is $\|\cdot\|_\infty$-Lipschitz with constant $\frac{2}{\rho}$. □

In the following we present the proof of Lemma 5.2.

Proof of Lemma 5.2. First consider the ℓ_∞-norm on the set $\mathcal{B}$. By Lemma 5.1, we have the function $\delta \mapsto \ell_\rho(W(x+\delta)y)$ is $\|\cdot\|_\infty$-Lipschitz with constant $\frac{2\Lambda_1}{\rho}$. Consider the set $C_{\mathcal{B}}(\epsilon\rho/4\Lambda_1)$. By Lemma B.1, and that $\|\delta\|_\infty \leq \beta$, for all $\delta \in \mathcal{B}$, we have

$$|C_{\mathcal{B}}(\epsilon\rho/4\Lambda_1)| \leq \left(\frac{12\Lambda_1\beta}{\epsilon\rho}\right)^d.$$

Therefore,

$$|\tilde{S}| = n\left(\frac{12\Lambda_1\beta}{\epsilon\rho}\right)^d,$$

and for $(\tilde{x}, y) \in \tilde{S}$ with $\tilde{x} = (x, \tilde{\delta})$,

$$\|\tilde{x}\|_2 \leq \|x\|_2 + \|\tilde{\delta}\|_2 \leq \Psi + \sqrt{d}\|\delta\|_\infty = \Psi'.$$

Thus, the result follows from Corollary B.4. □

In the following, we present the proof of Corollary 5.3.

Proof of Corollary 5.3. The proof is a direct application of Theorem 4.8 by setting α to $\frac{1}{n}$. Thus, consider the following integral

$$\begin{aligned}
\int_\alpha^1 \sqrt{\log \mathcal{N}_\infty(\epsilon/2, \tilde{\mathcal{G}}_{\text{adv}}, \tilde{S})}d\epsilon &\leq \int_{\frac{1}{n}}^1 \sqrt{C\frac{\Lambda^2(\Psi+\sqrt{d}\beta)^2}{\epsilon^2\rho^2}L_{\log}}d\epsilon \\
&\leq \sqrt{C}\frac{\tilde{L}_{\log}}{\log(n)}\frac{\Lambda\Psi'}{\rho}\int_{\frac{1}{n}}^1 \frac{1}{\epsilon}d\epsilon \leq \sqrt{C}\frac{\tilde{L}_{\log}}{\log(n)}\frac{\Lambda\Psi'}{\rho}\left[\log(\epsilon)\right]_{\frac{1}{n}}^1 \\
&= \sqrt{C}\frac{\tilde{L}_{\log}}{\log(n)}\frac{\Lambda\Psi'}{\rho}\log(n) = \sqrt{C}\tilde{L}_{\log}\frac{\Lambda\Psi'}{\rho}
\end{aligned}$$

The first inequality is due to the monotone property of integral. The second is by noticing that replacing ϵ by $\frac{1}{n}$ in $L_{\log}$ can only increase its value. Plugging this in Theorem 4.8 gives the result. □

B.2. Spatial Adversarial Attack

In this section we present the proof of adversarial generalization bound in Corollary 5.9. The first step is to show that the transformation $\delta \mapsto \ell(W(A(x,\delta), y)$ is $\|\cdot\|_\infty$-Lipschitz as summarized in Lemma B.5. We first show that $\delta \mapsto A(x,\delta)$ is $(\|\cdot\|_\infty, \|\cdot\|_\infty)$-Lipschitz. The following lemma states the result.

Lemma B.5. *Let $A(x,\delta)$ be defined as in (9). For all $\|x\|_1 \leq \Psi_1$, and $\delta, \delta' \in \mathcal{B}$, we have,*

$$\|A(x,\delta) - A(x,\delta')\|_\infty \leq 4\sqrt{d}\Psi_1\|\delta - \delta'\|_\infty.$$

Proof. Let U^s and V^s be the indexes corresponding to δ and $U^{s'}$ and $V^{s'}$ be indexes corresponding to δ'. Further let $a = \|(U^s, V^s) - (U^{s'}, V^{s'})\|_\infty$. Then by the definition of the ℓ_∞-norm and the transformation A, we have,

$$\begin{aligned}
&\|A(x,\delta) - A(x,\delta')\|_\infty \\
&= \max_{i\in[d]}\left|\sum_{k=1}^{\sqrt{d}}\sum_{l=1}^{\sqrt{d}} x_{kl}(\max(0, 1-|V_i^s - k|)\max(0, 1-|U_i^s - l|) - \max(0, 1-|V_i^{s'} - k|)\max(0, 1-|U_i^{s'} - l|))\right| \\
&\leq \max_{i\in[d]}\sum_{k=1}^{\sqrt{d}}\sum_{l=1}^{\sqrt{d}} |x_{kl}|\left|(\max(0, 1-|V_i^s - k|)\max(0, 1-|U_i^s - l|) - \max(0, 1-|V_i^{s'} - k|)\max(0, 1-|U_i^{s'} - l|))\right|,
\end{aligned}$$

where the inequality follows from the triangular inequality. For arbitrary $k, l, i \in [\sqrt{d}]$, let $a = \max(0, 1 - |V_i^s - k|)$, $b = \max(0, 1 - |U_i^s - l|)$, $a' = \max(0, 1 - |V_i^{s\prime} - l|)$, and $b' = \max(0, 1 - |U_i^{s\prime} - l|)$. Consider the following inequality

$$\begin{aligned}|ab - a'b'| &\leq |ab - ab'| + |ab' - a'b'| \\ &\leq |a||b - b'| + |b'||a - a'| \\ &\leq |b - b'| + |a - a'|.\end{aligned}$$

The first inequality is due to the triangular inequality and the last follows by $|a|, |b'| \leq 1$. Thus,

$$\begin{aligned}&\|A(x,\delta) - A(x,\delta')\|_\infty \\ &\leq \max_{i\in[d]} \sum_{k=1}^{\sqrt{d}} \sum_{l=1}^{\sqrt{d}} |x_{kl}| \left(|\max(0, 1 - |U_i^s - k|) - \max(0, 1 - |U_i^{s\prime} - k)|\right| \\ &\qquad + |\max(0, 1 - |V_i^s - k|) - \max(0, 1 - |V_i^{s\prime} - k)|) \\ &\leq \max_{i\in[d]} \sum_{k=1}^{\sqrt{d}} \sum_{l=1}^{\sqrt{d}} |x_{kl}| \left(||U_i^s - k| - |U_i^{s\prime} - k|| + ||V_i^s - k| - |V_i^{s\prime} - k||\right) \\ &\leq \max_{i\in[d]} \sum_{k=1}^{\sqrt{d}} \sum_{l=1}^{\sqrt{d}} |x_{kl}| \left(|U_i^s - U_i^{s\prime}| + |V_i^s - V_i^{s\prime}|\right) \\ &\leq \sum_{k=1}^{\sqrt{d}} \sum_{l=1}^{\sqrt{d}} |x_{kl}|\, 2\|(U^s, V^s) - (U^{s\prime}, V^{s\prime})\|_\infty \leq 2\Psi_1 \|(U^s, V^s) - (U^{s\prime}, V^{s\prime})\|_\infty.\end{aligned}$$

The first inequality is due to the triangular inequality. The second inequality follows from 1-Lipschitzness of the function $x \mapsto \max(0, x)$. The third inequality is due to the reverse triangular inequality (i.e., $||c| - |d|| \leq |c - d|$). Now it remains to derive a bound for $\|(U^s, V^s) - (U^{s\prime}, V^{s\prime})\|_\infty$ by $\|\delta - \delta'\|_\infty$. Observe the following

$$\begin{aligned}\|(U^s, V^s) - (U^{s\prime}, V^{s\prime})\|_\infty &= \max_{i\in[d]} \max(|V_i^s - V_i^{s\prime}|, |U_i^s - U_i^{s\prime}|) \\ &= \max_{i\in[d]} \max(|V_i^t(\delta_{22} - \delta'_{22}) - U_i^t(\delta_{21} - \delta'_{21})|, |V_i^t(\delta_{12} - \delta'_{12}) - U_i^t(\delta_{11} - \delta'_{11})|) \\ &\leq \max_{i\in[d]} \max(|V_i^t||\delta_{22} - \delta'_{22}| + |U_i^t||\delta_{21} - \delta'_{21}|, |V_i^t||\delta_{12} - \delta'_{12}| + |U_i^t||\delta_{11} - \delta'_{11})) \\ &\leq 2\sqrt{d}\|\delta - \delta'\|_\infty.\end{aligned}$$

The second equality is due to the definition of $\mathcal{S}$ and (7). The first inequality follows from the triangular inequality. The last inequality is derived from the fact $\max_{i\in[d]} |V_i^t| = \sqrt{d}$. Combining the two inequalities we get

$$\|A(x,\delta) - A(x,\delta')\|_\infty \leq 4\sqrt{d}\Psi_1 \|\delta - \delta'\|_\infty.$$

The proof is completed. □

We now utilize this result to prove Lemma 5.7.

Proof of Lemma 5.7. For $\delta, \delta' \in \mathcal{B}$, we then have

$$\begin{aligned}|\ell_\rho(WA(x,\delta), y) - \ell_\rho(WA(x,\delta'), y)| &\leq \frac{2}{\rho}\|WA(x,\delta) - WA(x,\delta')\|_\infty \\ &\leq \frac{2}{\rho} \max_i |\langle W_{i,\cdot}, (A(x,\delta) - A(x,\delta'))\rangle| \\ &\leq \frac{2}{\rho} \max_i \|W_{i,\cdot}\|_1 \|A(x,\delta) - A(x,\delta')\|_\infty \\ &\leq \frac{8}{\rho}\Lambda_1\Psi_1\sqrt{d}\|\delta - \delta'\|_\infty,\end{aligned}$$

where the first inequality follows from the Lipschitzness of the loss ℓ_ρ, the second from the definition of ℓ_∞-norm, the third from the Hölder inequality, and the final from Lemma B.5. □

Proof of Lemma 5.8. We consider the ℓ_∞-covering number of the set $\mathcal{B}$. Recall that, we have, by Lemma 5.7, the function $\delta \mapsto \ell_\rho(WA(x,\delta),y)$, is $\|\cdot\|_\infty$-Lipschitz with the Lipschitz constant $\frac{8}{\rho}\Lambda_1\Psi_1\sqrt{d}$. We now aim to control the size of the set $C_{\mathcal{B}}(\epsilon\rho/16\Lambda_1\Psi_1\sqrt{d})$. By Lemma B.1, for the ball $\mathcal{B} = \{\delta : \|\delta\|_\infty \le \beta\}$, we have

$$\left|C_{\mathcal{B}}(\epsilon\rho/16\Lambda_1\Psi_1\sqrt{d})\right| \le \left(\frac{48\Lambda_1\Psi_1\sqrt{d}\beta}{\epsilon\rho}\right)^4.$$

By the construction of $\tilde{S}$, we then have

$$|\tilde{S}| = n\left(\frac{48\Lambda_1\Psi_1\sqrt{d}(\beta}{\epsilon\rho}\right)^4.$$

Further note that $\|A(x,\delta)\|_2 \le 4\|x\|_2 \le 4\Psi$, for all $x \in \mathcal{X}$, since spatial transformation does not alter the norm of the input except for the factor of 4 due to the bilinear interpolation. Thus, the result then follows from Corollary B.4.

□

C. Proofs of Applications to Neural Networks in Section 5.2

In this section, we present the omitted proofs of section 5.2. The technique is similar to the linear case. We first establish the Lipschitzness property of the functions $\delta \mapsto \ell_\rho(N_{\mathcal{W}}(A(x,\delta)),y)$. We then extend the data set and apply ℓ_∞-covering number results of the neural networks function class.

We first review the following Lemma (Ledent et al., 2021b). It establishes a bound on the ℓ_∞-covering numbers of norm-bounded neural network function classes.

Lemma C.1 (Ledent et al. 2021b)**.** *Let $\mathcal{F}$ be the class of neural networks that is, $\mathcal{F} = \{x \mapsto N_{\mathcal{W}}(x)\}$, where $\mathcal{W} = (W^1,\dots,W^L)$ are a set of weights and $N_{\mathcal{W}} = W^L\sigma(W^{L-1}\sigma(\dots W^1x))$ with 1-Lipschitz element-wise non-linearities σ. Define the loss class $\mathcal{L} = \ell_p \circ \mathcal{F}$ where ℓ_p is defined as (5). Suppose that $\|W^l\|_{21} \le a_l$ and $\|W^l\|_\sigma \le s_l$ for all $l \in [L-1]$, $\|W^L\|_2 \le a_L$, $\|W^L\|_{2,\infty} \le s_L$, $\|x\|_2 \le b$, and m_l is the width of the l'th layer. Then given a data set S with n elements and $\epsilon > 0$,*

$$\log \mathcal{N}_\infty(\epsilon,\mathcal{L},S) \le \frac{CL^2b^2}{\rho^2\epsilon^2}\prod_{l=1}^L s_l^2\left(\sum_{l=1}^L \frac{a_l^2}{s_l^2}\right)\log\left((\,C_1b\Gamma/(\epsilon\rho) + C_2\bar{m}\,)n+1\right),$$

where $\Gamma = \max_{l\in[L]}(\prod_{i=1}^L s_i)a_lm_l/s_l$, $\bar{m} = \max_{l\in[L]} m_l$, and C, C_1, C_2 are universal constants.

C.1. ℓ_∞-additive Perturbation Attack

The first step is to derive the $\|\cdot\|_\infty$-Lipschitz constant of the loss function as a function in the noise parameter. We start by the following lemma on $\|\cdot\|_\infty$-Lipschitzness of neural network as a function of the input.

Lemma C.2. *Consider the neural network function $N_{\mathcal{W}}(x)$ defined as (10). Given $x, x' \in \mathcal{X}$, then for all $y \in \mathcal{Y}$*

$$|\ell_\rho(N_{\mathcal{W}}(x),y) - \ell_\rho(N_{\mathcal{W}}(x'),y)| \le \frac{2}{\rho}\left(\prod_{l=2}^L s_l\right)\sqrt{m_1}s_1'\|x-x'\|_\infty.$$

Proof. Let $N^l_{\mathcal{W}}$ be the output of the l'th layer of the network $N_{\mathcal{W}}$. Consider the following,

$$\begin{aligned}
|\ell_\rho(N_{\mathcal{W}}(x),y)-\ell_\rho(N_{\mathcal{W}}(x'),y)| &\le \frac{2}{\rho}\|N_{\mathcal{W}}(x)-N_{\mathcal{W}}(x')\|_\infty \\
&= \frac{2}{\rho}\|W^L(N^{L-1}_{\mathcal{W}}(x)-N^{L-1}_{\mathcal{W}}(x'))\|_\infty \\
&\le \frac{2}{\rho}\max_{i\in[m_L]}\|W^L_{i,.}\|_2\|(N^{L-1}_{\mathcal{W}}(x)-N^{L-1}_{\mathcal{W}}(x'))\|_2 \\
&\le \frac{2}{\rho}\prod_{l=2}^{L}s_l\|W^1x-W^1x'\|_2 \le \frac{2}{\rho}\prod_{l=2}^{L}s_l\sqrt{m_1}\|W^1x-W^1x'\|_\infty \\
&\le \frac{2}{\rho}\prod_{l=2}^{L}s_l\sqrt{m_1}\max_{i\in[m_1]}\|W^1_{i,.}\|_1\|x-x'\|_\infty = \frac{2}{\rho}\prod_{l=2}^{L}s_l\sqrt{m_1}s'_1\|x-x'\|_\infty.
\end{aligned}$$

The second inequality follows from the definition of ℓ_∞-norm and Hölder inequality, the third from the fact that the non-linearity is 1-Lipschitz and by induction over the layers. The fourth is from the fact that $\|x\|_2 \le \sqrt{d}\|x\|_\infty$, for all $x \in \mathbb{R}^d$. □

We now present the proof of Lemma 5.13.

Proof of Lemma 5.13. The result follows directly from Lemma C.2. Let $\delta, \delta' \in \mathcal{B}$, then for all $(x,y) \in \mathcal{Z}$

$$|\ell_\rho(N_{\mathcal{W}}(x+\delta),y)-\ell_\rho(N_{\mathcal{W}}(x+\delta'),y)| \le \frac{2}{\rho}\prod_{l=2}^{L}s_l\sqrt{m_1}s'_1\|(x+\delta)-(x+\delta')\|_\infty = \frac{2}{\rho}\prod_{l=2}^{L}s_l\sqrt{m_1}s'_1\|\delta-\delta'\|_\infty.$$

The proof is completed. □

Proof of Lemma 5.14. We consider ℓ_∞ covering number of set $\mathcal{B}$. By Lemma 5.13, the function $\delta \mapsto \ell_\rho(N_{\mathcal{W}}(x+\delta),y)$, is $\|\cdot\|_\infty$-Lipschitz with the Lipschitz constant $\frac{2}{\rho}\prod_{l=2}^{L}s_l\sqrt{m_1}s'_1$. By Lemma 4.4, we aim now to bound the size of the set $C_{\mathcal{B}}(\epsilon\rho/4\prod_{l=2}^{L}s_l\sqrt{m_1}s'_1)$. By Lemma B.1, and that $\|\delta\|_\infty \le \beta$, we have

$$\left|C_{\mathcal{B}}\left(\epsilon\rho/4\prod_{l=2}^{L}s_l\sqrt{m_1}s'_1\right)\right| \le \left(\frac{12\prod_{l=2}^{L}s_l\sqrt{m_1}s'_1\beta}{\epsilon\rho}\right)^d.$$

Therefore,

$$|\tilde{S}| = n\left(\frac{12\prod_{l=2}^{L}s_l\sqrt{m_1}s'_1\beta}{\epsilon\rho}\right)^d,$$

and for $(\tilde{x},y) \in \tilde{S}$ with $\tilde{x} = (x,\tilde{\delta})$, we have

$$\|\tilde{x}\|_2 \le \|x\|_2 + \|\tilde{\delta}\|_2 \le \Psi + \sqrt{d}\|\delta\|_\infty = \Psi'.$$

Thus, the result follows from Lemma C.1. □

C.2. Spatial Adversarial Attack

Proof of Lemma 5.18. The proof follows directly from Lemmas C.2 and B.5. Let $\delta, \delta' \in \mathcal{B}$, then for all $\mathcal{W} \in \mathbb{W}$, $(x,y) \in \mathcal{Z}$,

$$\begin{aligned}
|\ell_\rho(N_{\mathcal{W}}(A(x,\delta)),y)-\ell_\rho(N_{\mathcal{W}}(A(x,\delta')),y)| &\le \frac{2}{\rho}\prod_{l=2}^{L}s_l\sqrt{m_1}s'_1\|A(x,\delta))-A(x,\delta'))\|_\infty \\
&\le \frac{8}{\rho}\prod_{l=2}^{L}s_l\sqrt{m_1}s'_1\Psi_1\sqrt{d}\|\delta-\delta'\|_\infty.
\end{aligned}$$

The proof is completed. □

Now we present the proof of Lemma 5.19.

Proof of Lemma 5.19. The function $\delta \mapsto \ell_\rho(N_{\mathcal{W}}(A(x,\delta)),y)$ is $\|\cdot\|_\infty$-Lipschitz with the Lipschitz constant $\frac{8}{\rho}\prod_{l=2}^{L} s_l\sqrt{m_1}s_1'\Psi_1\sqrt{d}$. By Lemma 4.4, we aim now to bound the size of the set $C_{\mathcal{B}}(\epsilon\rho/16\prod_{l=2}^{L} s_l\sqrt{m_1}s_1'\Psi_1\sqrt{d})$. By Lemma B.1, and that $\|\delta\|_\infty \leq \beta$, we have

$$\left|C_{\mathcal{B}}\left(\epsilon\rho/16\prod_{l=2}^{L} s_l\sqrt{m_1}s_1'\Psi_1\sqrt{d}\right)\right| \leq \left(\frac{48\prod_{l=2}^{L} s_l\sqrt{m_1}s_1'\Psi_1\sqrt{d})\beta}{\epsilon\rho}\right)^4.$$

Therefore, we have

$$|\tilde{S}| = n\left(\frac{48\prod_{l=2}^{L} s_l\sqrt{m_1}s_1'\Psi_1\sqrt{d}\beta}{\epsilon\rho}\right)^4.$$

We again note that $\|A(x,\delta)\|_2 \leq 4\|x\|_2$, by the fact that spatial transformation does not alter the input norms except for the factor of 4 from the bilinear interpolation. The result, thus, follows by Lemma C.1.

□

D. Proofs of Optimistic Bounds

Proof of Lemma 6.5. The proof strategy is to show that the covering number of the set $\mathcal{G}_{\text{adv}}|_S^r$ is controlled by the covering number of a set of functions of the adversarial noise. To that extent define the set $\mathcal{H}_S|_r =$

$$\left\{\left(f(A(x_1,.),w)_1,\ldots,f(A(x_1,.),w)_K,\ldots f(A(x_n,.),w)_1,\ldots f(A(x_n,.),w)_K\right), w\in\mathbb{W}, \hat{R}_{\text{adv}}(w)\leq r\right\} \subset (\mathbb{R}^{nK})^{\mathcal{B}}$$

and

$$\mathcal{H}_S = \left\{\left(f(A(x_1,.),w)_1,\ldots,f(A(x_1,.),w)_K,\ldots f(A(x_n,.),w)_1,\ldots f(A(x_n,.),w)_K\right), w\in\mathbb{W}\right\} \subset (\mathbb{R}^{nK})^{\mathcal{B}}.$$

Let

$$C_{\mathcal{H}_S|_r} = \left\{\left(c_1^j(\cdot)_1,\ldots,c_1^j(\cdot)_K,\ldots c_n^j(\cdot)_1,\ldots c_n^j(\cdot)_K\right), w\in\mathbb{W}\right\} \subset (\mathbb{R}^{nK})^{\mathcal{B}}$$

be an $(\frac{\epsilon}{\lambda(2r)^\theta},\ell_\infty)$-cover of $\mathcal{H}_S|_r$. Further suppose that it is a proper cover, that is $C_{\mathcal{H}_S|_r}\subset\mathcal{H}_S|_r$. Note that for any $w\in\mathbb{W}$ there exists a $j(w)$ such that

$$\max_{i\in[n]}\max_{\delta\in\mathcal{B}}\max_{k\in[K]}|f(A(x_i,\delta),w) - c_i^{j(w)}(\delta)_k| \leq \frac{\epsilon}{\lambda(2r)^\theta}.$$

We now claim that $C_{\mathcal{H}_S|_r}$ can be used to cover the set $\mathcal{G}|_S^r$ at an ϵ resolution as measured by ℓ_2 norm. In other words, we aim to show that

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\max_{\delta\in\mathcal{B}}\ell(f(A(x_i,\delta),w),y_i) - \max_{\delta\in\mathcal{B}}\ell(c_i^{j(w)}(\delta),y_i)\right)^2} \leq \epsilon.$$

Observe the following

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\max_{\delta\in\mathcal{B}}\ell(f(A(x_i,\delta),w),y_i)-\max_{\delta\in\mathcal{B}}\ell(c_i^{j(w)}(\delta),y_i)\right)^2\\
\leq&\frac{1}{n}\sum_{i=1}^{n}\lambda^2\max\{\max_{\delta\in\mathcal{B}}\ell(f(A(x_i,\delta),w),y_i),\max_{\delta\in\mathcal{B}}\ell(c_i^{j(w)}(\delta),y_i)\}^{2\theta}\max_{\delta\in\mathcal{B}}\|f(A(x_i,\delta),w)-c_i^{j(w)}(\delta)\|_\infty^2\\
\leq&\frac{1}{n}\sum_{i=1}^{n}\lambda^2\max\{\max_{\delta\in\mathcal{B}}\ell(f(A(x_i,\delta),w),y_i),\max_{\delta\in\mathcal{B}}\ell(c_i^{j(w)}(\delta),y_i)\}^{2\theta}\max_{k\in[K]}\max_{i\in[n]}\max_{\delta\in\mathcal{B}}\|f(A(x_i,\delta),w)_k-c_i^{j(w)}(\delta)_k\|_\infty^2\\
=&\left(\frac{\epsilon\lambda}{\lambda(2r)^\theta}\right)^2\frac{1}{n}\sum_{i=1}^{n}\max\{\max_{\delta\in\mathcal{B}}\ell(f(A(x_i,\delta),w),y_i),\max_{\delta\in\mathcal{B}}\ell(c_i^{j(w)}(\delta),y_i))\}^{2\theta}\\
\leq&\left(\frac{\epsilon\lambda}{\lambda(2r)^\theta}\right)^2\frac{1}{n}\sum_{i=1}^{n}(\max_{\delta\in\mathcal{B}}\ell(f(A(x_i,\delta),w),y_i)+\max_{\delta\in\mathcal{B}}\ell(c_i^{j(w)}(\delta),y_i))^{2\theta}\\
\leq&\left(\frac{\epsilon\lambda}{\lambda(2r)^\theta}\right)^2\left(\frac{1}{n}\sum_{i=1}^{n}\max_{\delta\in\mathcal{B}}\ell(f(A(x_i,\delta),w),y_i)+\frac{1}{n}\sum_{i=1}^{n}\max_{\delta\in\mathcal{B}}\ell(c_i^{j(w)}(\delta),y_i)\right)^{2\theta}\\
\leq&\left(\frac{\epsilon\lambda}{\lambda(2r)^\theta}\right)^2(2r)^{2\theta}=\epsilon^2.
\end{aligned}
$$

Since we can construct a proper cover at precision ϵ from a general cover at precision $\epsilon/2$, we conclude that

$$\mathcal{N}_2(\epsilon,\mathcal{G}_{adv}|^r,S)=\mathcal{N}_2(\epsilon,\mathcal{G}_{adv}|_S^r)\leq\mathcal{N}_\infty(\frac{\epsilon}{2\lambda(2r)^\theta},\mathcal{H}_S|_r).$$

Furthermore, since $\mathcal{H}_S|_r\subset\mathcal{H}_S$, we have

$$\mathcal{N}_2(\epsilon,\mathcal{G}_{adv}|^r,S)\leq\mathcal{N}_\infty(\frac{\epsilon}{2\lambda(2r)^\theta},\mathcal{H}_S|_r)\leq\mathcal{N}_\infty(\frac{\epsilon}{2\lambda(2r)^\theta},\mathcal{H}_S).$$

The following step is to show that we can control $\mathcal{N}_\infty(\frac{\epsilon}{\lambda(2r)^\theta},\mathcal{H}_S)$ by a cover of a discretized version of $\mathcal{H}_S$. This holds by Lemma 4.4 and hence completes the proof. □

Proof of Lemma 6.6. Note that by the definition of M_ϵ, we have $|\hat{S}|=nKM_\epsilon$.

Now our goal is to bound the local Rademacher complexity of the adversarial class $\mathfrak{R}_n(\mathcal{F}_{adv}|_r)$. By Theorem A.2 we have

$$\mathfrak{R}(\mathcal{G}_{adv}|^r)\leq\inf_{\alpha>0}\left(4\alpha+\frac{10}{\sqrt{n}}\int_\alpha^{\sqrt{br}}\sqrt{\log\mathcal{N}_2(\epsilon,\mathcal{G}_{adv}|^r,S)}d\epsilon\right).$$

Select $\alpha=4\lambda(2r)^\theta/\sqrt{n}$ then

$$\mathfrak{R}(\mathcal{G}_{adv}|^r)\leq16\lambda(2r)^\theta/\sqrt{n}+\frac{10}{\sqrt{n}}\int_{4\lambda(2r)^\theta/\sqrt{n}}^{\sqrt{br}}\sqrt{\log\mathcal{N}_\infty(\frac{\epsilon}{4\lambda(2r)^\theta},\tilde{\mathcal{F}}_{\mathrm{adv}},\hat{S})}d\epsilon.$$

Now by change of variable we get

$$
\begin{aligned}
\mathfrak{R}(\mathcal{G}_{\mathrm{adv}}|^r)&\leq16\lambda(2r)^\theta/\sqrt{n}+\frac{40\lambda(2r)^\theta}{\sqrt{n}}\int_{\frac{1}{\sqrt{n}}}^{\frac{b^{1/2}r^{1/2-\theta}}{2^{2+\theta}\lambda}}\sqrt{\log\mathcal{N}_\infty(\epsilon,\tilde{\mathcal{F}},\hat{S})}d\epsilon\\
&\leq16\lambda(2r)^\theta/\sqrt{n}+\frac{40\lambda(2r)^\theta}{\sqrt{n}}\int_{\frac{1}{\sqrt{n}}}^{\frac{b^{1-\theta}}{2^{2+\theta}\lambda}}\sqrt{\log\mathcal{N}_\infty(\epsilon,\tilde{\mathcal{F}}_{\mathrm{adv}},\hat{S})}d\epsilon,
\end{aligned}
$$

where the second inequality is due to the fact that $b \geq r$. By the assumption $\mathcal{N}_\infty\left(\epsilon, \tilde{\mathcal{F}}_{\text{adv}}, \hat{S}\right) \leq \frac{R_a}{\epsilon^2}$, for $\epsilon \in [a, b]$, we have

$$\begin{aligned}
\mathfrak{R}(\mathcal{G}_{adv}|^r) &\leq 16\lambda(2r)^\theta/\sqrt{n} + \frac{40\lambda(2r)^\theta}{\sqrt{n}} \int_{\frac{1}{\sqrt{n}}}^{\frac{b^{1-\theta}}{2^{2+\theta}\lambda}} \sqrt{\log \mathcal{N}_\infty(\epsilon, \tilde{\mathcal{F}}_{\text{adv}}, \hat{S})} d\epsilon \\
&\leq 16\lambda(2r)^\theta/\sqrt{n} + \frac{40\lambda(2r)^\theta}{\sqrt{n}} \int_{\frac{1}{\sqrt{n}}}^{\frac{b^{1-\theta}}{2^{2+\theta}\lambda}} \frac{\sqrt{R_{\frac{1}{\sqrt{n}}}}}{\epsilon} d\epsilon \\
&= 16\lambda(2r)^\theta/\sqrt{n} + \frac{40\lambda(2r)^\theta \sqrt{R_{\frac{1}{\sqrt{n}}}}}{\sqrt{n}} \int_{\frac{1}{\sqrt{n}}}^{\frac{b^{1-\theta}}{2^{2+\theta}\lambda}} \frac{1}{\epsilon} d\epsilon \\
&= 16\lambda(2r)^\theta/\sqrt{n} + \frac{40\lambda(2r)^\theta \sqrt{R_{\frac{1}{\sqrt{n}}}}}{\sqrt{n}} \left[\log(\frac{b^{1-\theta}}{2^{2+\theta}\lambda}) - \log(\frac{1}{\sqrt{n}})\right] \\
&= \lambda r^\theta/\sqrt{n} \left[2^{4+\theta} + 40\sqrt{R_{\frac{1}{\sqrt{n}}}} \log\left(\frac{b^{1-\theta}\sqrt{n}}{2^{2+\theta}\lambda}\right)\right].
\end{aligned}$$

The proof is completed. □

To prove Theorem 6.8 we require the following lemma, which gives optimistic bounds for learning with smooth loss functions. We say ϕ is sub-root if $r \mapsto \phi(r)/\sqrt{r}$ is nonincreasing. We say r^* is a fixed-point of ϕ if $r^* = \phi(r^*)$.

Lemma D.1 (Srebro et al. 2010)**.** *Consider a real hypothesis class $\mathcal{F}$. Further assume that, for all $f \in \mathcal{F}$, $\|f(x)\| \leq B$. Let ℓ be a loss function bounded by b. Let S be sampled i.i.d. from some distribution $\mathcal{D}$. Suppose that the local Rademacher complexity $\mathfrak{R}_n(\ell \circ \mathcal{F}|^r) \leq \phi(r)$, where ϕ is a sub-root functions. We have with probability at least $1-\delta$, for all $f \in \mathcal{F}$,*

$$R(f) \leq \hat{R}(f) + 106r^* + \frac{48b}{n}(\log(1/\delta) + \log(\log(n))) + \sqrt{\hat{R}(f)\left(8r^* + \frac{4b}{n}(\log(1/\delta) + \log(\log(n)))\right)}.$$

Proof of Theorem 6.8. Note that by Lemma D.1, with probability at least $1-\delta$ we have the following inequality for all $w \in \mathbb{W}$,

$$R_{\text{adv}}(w) \leq \hat{R}_{\text{adv}}(w) + 106r^* + \frac{48b}{n}(\log(1/\delta) + \log(\log(n))) + \sqrt{\hat{R}_{\text{adv}}(w)\left(8r^* + \frac{4b}{n}(\log(1/\delta) + \log(\log(n)))\right)},$$

where r^* is the fixed point solution of the sub-root function $\phi(r)$ satisfying $\mathfrak{R}_n(\mathcal{G}_{\text{adv}}) \leq \phi(r)$. Observe also that by Lemma 6.6 and that $\theta = 1/2$ a good candidate of ϕ is $\phi(r) = \lambda\sqrt{2r}/\sqrt{n}\left[16 + 40\sqrt{R_{\frac{1}{\sqrt{n}}}} \log\left(\frac{b^{1-\theta}\sqrt{n}}{2^{2+\theta}\lambda}\right)\right]$. Therefore, $r^* = \frac{\lambda^2}{n}\left[16\sqrt{2} + 40\sqrt{R_{\frac{1}{\sqrt{n}}}} \log\left(\frac{b^{1-\theta}\sqrt{n}}{2^{2+\theta}\lambda}\right)\right]^2$. It follows the following inequality with probability at least $1-\delta$ for all w

$$\begin{aligned}
R_{\text{adv}}(w) \leq& \hat{R}_{\text{adv}}(w) + \frac{106\lambda^2}{n}\left[16\sqrt{2} + 40\sqrt{R_{\frac{1}{\sqrt{n}}}} \log\left(\frac{b^{1-\theta}\sqrt{n}}{2^{2+\theta}\lambda}\right)\right]^2 + \frac{48b}{n}(\log(1/\delta) \\
&+ \log(\log(n))) + \sqrt{\hat{R}_{\text{adv}}(w)\left(\frac{8\lambda^2}{n}\left[16\sqrt{2} + 40\sqrt{R_{\frac{1}{\sqrt{n}}}} \log\left(\frac{b^{1-\theta}\sqrt{n}}{2^{2+\theta}\lambda}\right)\right]^2 + \frac{4b}{n}(\log(1/\epsilon) + \log(\log(n)))\right)}.
\end{aligned}$$

The proof is completed. □

E. Example of Robust-self-bounding Loss

In this section we show that the loss function

$$L_\rho(t, y) = \begin{cases} 1 & \text{if } M(t,y) \leq 0, \\ 2\left(M(t,y)/\rho\right)^3 - 3\left(M(t,y)/\rho\right)^2 + 1 & \text{if } 0 < M(t,y) < \rho, \\ 0 & \text{if } M(t,y) \geq \rho, \end{cases}$$

is a robust-self-bounding function. It was shown in Reeve & Kaban (2020) that for $t, t' \in \mathbb{R}^K$

$$|L_\rho(t, y) - \ell(t', y)| \leq 2\sqrt{6}/\rho \max\{\ell(t, y), \ell(t', y)\}^{\frac{1}{2}} \|t - t'\|_\infty. \tag{16}$$

Now consider the function $\nu : \mathcal{B} \to \mathbb{R}^K$ and $\mu : \mathcal{B} \to \mathbb{R}^K$. Then

$$\begin{aligned}
|\max_{\delta\in\mathcal{B}} L_\rho(\mu(\delta), y) - \max_{\delta\in\mathcal{B}} L_\rho(\nu(\delta), y)| &\leq \max_{\delta\in\mathcal{B}} |L_\rho(\mu(\delta), y) - L_\rho(\nu(\delta), y)| \\
&\leq \max_{\delta\in\mathcal{B}} 2\sqrt{6}/\rho \max\{L_\rho(\mu(\delta), y), L_\rho(\nu(\delta), y)\}^{\frac{1}{2}} \|\mu(\delta) - \nu(\delta)\|_\infty \\
&\leq 2\sqrt{6}/\rho \max\{\max_{\delta\in\mathcal{B}} L_\rho(\mu(\sigma), y), \max_{\delta\in\mathcal{B}} L_\rho(\nu(\sigma), y)\}^{\frac{1}{2}} \max_{\delta\in\mathcal{B}} \|\mu(\delta) - \nu(\delta)\|_\infty,
\end{aligned}$$

where the second inequality is due to Eq. (16).