Why Clean Generalization and Robust Memorization both Happen in Adversarial Training

Anonymous Authors[1]

Abstract

Adversarial training is a standard method to train neural networks to be robust to adversarial perturbation. However, in contrast with benign overfitting in the standard deep learning setting, which means that over-parameterized neural networks surprisingly generalize well for unseen data, while adversarial training method is able to achieve low robust training error, there still exists a significant robust generalization gap, which promotes us exploring what mechanism leads to robust overfitting during learning process. In this paper, we will study a phenomenon called **Clean Generalization and Robust Memorization (CGRM)** in adversarial training under the realistic data assumption. By function approximation theory, we prove that ReLU nets with efficient size have the ability to achieve CGRM, while robust generalization requires exponentially large models. Then, we demonstrate CGRM in adversarial training from both empirical and theoretical perspectives. In particular, we empirically investigate the dynamics of loss landscape over input, and we also provide theoretical analysis of CGRM on data with linear separable assumption. Finally, we prove novel generalization bounds, which further explains why deep neural networks have both high clean test accuracy and robust overfitting in adversarial training.

1. Introduction

Although deep learning has made a remarkable success in many application fields, such as computer vision (Voulodimos et al., 2018) and natural language process (Devlin et al., 2018), it is well-known that only small but adversarial perturbations additional to the natural data can make well-trained classifiers confused (Szegedy et al., 2013; Goodfellow et al., 2014), which promotes designing adversarial robust learning algorithms. In practice, adversarial training methods (Madry et al., 2017; Shafahi et al., 2019; Zhang et al., 2019) are widely used to improve the robustness of models by regarding perturbed data as training data. However, while these robust learning algorithms are able to achieve high robust training accuracy, it still leads to a non-negligible robust generalization gap (Raghunathan et al., 2019), which is also called *robust overfitting* (Rice et al., 2020; Yu et al., 2022).

To explain this puzzling phenomenon, a series of works have attempted to provide theoretical understandings from different perspectives. Despite these aforementioned works seem to provide a series of convincing evidence from theoretical views in different settings, there still exists a gap between theory and practice for at least two reasons.

First, although previous works have shown that adversarial robust generalization requires more data and larger models (Schmidt et al., 2018; Li et al., 2022), it is unclear that what mechanism, during adversarial training process, directly causes robust overfitting. In other words, we know there is no robust generalization gap for a trivial model that only guesses labels totally randomly (e.g. deep neural networks at random initialization), which implies that we should take learning process into consideration to analyze robust generalization.

Second and most importantly, while some works (Tsipras et al., 2018; Zhang et al., 2019) point out that achieving robustness may hurt clean test accuracy, in most of the cases, it is observed that drop of robust test accuracy is much higher than drop of clean test accuracy in adversarial training (Schmidt et al., 2018; Raghunathan et al., 2019) (see in Table 1). Namely, a weak version of benign overfitting, which means that overparameterized deep neural networks can both fit random data powerfully and generalize well for unseen clean data, remains after adversarial training.

Therefore, it is natural to ask the following question:

What happens, during adversarial learning process, resulting in ***Clean Generalization and Robust Memorization (CGRM)*** *at the same time?*

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Table 1. Train and test performances on CIFAR-10 (Raghunathan et al., 2019)

	Clean training	Adversarial training
Robust test	3.5%	45.8%
Robust train	–	100%
Clean test	95.2%	87.3%
Clean train	100%	100%

In this paper, we provide a theoretical understanding of the adversarial training process by proposing a novel implicit bias called *clean generalization and robust memorization (CGRM)* under the realistic data assumption, which explains why adversarial training leads to both high clean test accuracy and robust generalization gap.

The fundamental data structure that we leverage is that data can be clean separated by a neural network f_{clean} with moderate size but this neural classifier is non-robust on data with small margin, which is consistent with the common practice that well-trained neural networks have good clean performance but fail to classify perturbed data due to the close distance between small margin data and decision boundary. The existence of small and large margin data for image classification has been empirically verified (Banburski et al., 2021). And we also assume that data is well-separated, which means that there exists a robust classifier in general (Yang et al., 2020). However, this robust classifier may be hard to approximate by neural networks (Li et al., 2022). In other words, clean training often finds simple but non-robust solution, although robust classifier always exists.

Specifically, we consider the underlying data distribution $\mathcal{D}$ where μ proportional data is with small margin, which can be perturbed as adversarial example, and $1-\mu$ proportional data is with large margin, which can be robustly classified by the neural network f_{clean}. In adversarial training, we access N instances $\mathcal{S}$ randomly drawn from $\mathcal{D}$ as training data. In fact, by applying concentration technique in high dimensional probability, we know that there roughly exist μN small margin training data and $(1-\mu)N$ large margin training data.

In order to answer the question that we ask, under the above realistic data assumption, we consider the following classifier,

$$f_{\text{adv}}(x) := \underbrace{\sum_{(x_i,y_i)\in\mathcal{S}_{\text{small}}} y_i \mathbb{I}\{\|x-x_i\|_p \le \delta\}}_{\text{robust local indicators on small margin data}} + \underbrace{f_{\text{clean}}(x)\mathbb{I}\left\{\min_{(x_i,y_i)\in\mathcal{S}_{\text{small}}} \|x-x_i\|_p > \delta\right\}}_{\text{clean classifier on other data}},$$

where $\mathcal{S}_{\text{small}}$ is the small margin part of training dataset, and δ denotes adversarial perturbation radius, which is smaller than the separation between data in our setting.

Indeed, the classifer f_{adv} satisfies all main characteristics of deep models outputted by adversarial training (Proposition 3.2): *first*, adversarial training error with respect to f_{adv} achieves zero; *second*, f_{adv} has a good clean performance because the classifier is robust within the neighborhood of training data with small margin and it performs as the same as the clean classifier on other data; *third*, although this classifier has low robust training error, it is non-robust on other data since it only robustly memorizes training data with small margin, which results in robust overfitting.

Inspired by the ideal model f_{adv}, we then propose a conjecture that deep neural networks converge to solution similar to f_{adv} in adversarial training, which is a novel implicit bias of adversarial training, and we call it **clean generalization and robust memorization (CGRM)**, which provides a theoretical understanding of adversarial training, including why solution found by adversarial training has both good clean generalization and robust generalization gap.

Based on function approximation theory, we can prove that, for $d-$dimensional training dataset $\mathcal{S}$ with N samples, ReLU networks with $\widetilde{O}(\mu N d + \text{poly}(d))$ parameters are able to represent the target functions f_{adv} that we mention for CGRM (Theorem 3.3 and Corollary 3.4). However, we still prove a lower bound for the network size that is exponential in the data dimension d to make robust generalization (Theorem 3.5), and it manifests that robust classifier has higher representation complexity than the classifier with CGRM, which potentially explain why adversarial training converges to CGRM regime rather than finds robust classifier due to the inductive bias of low representation complexity.

To further demonstrate CGRM in adversarial training, we empirically investigate the dynamics of loss landscape over input during adversarial learning process. Concretely, for the loss function $\mathcal{L}(f(x), y)$ over classifier f and on data point (x, y), we verify whether $\hat{f}$ outputted by adversarial training is with CGRM by certain measure — maximum gradient norm within the neighborhood of training data input,

$$\frac{1}{N}\sum_{i=1}^{N} \max_{\|\xi\|_p \le \delta} \|\nabla_x \mathcal{L}(f(x_i+\xi), y_i)\|_q,$$

where $\{(x_i, y_i)\}_{i=1}^{N}$ denotes the training dataset, and $\frac{1}{p} + \frac{1}{q} = 1$.

This measure reflects the local flatness of loss landscape over input on training dataset. In CGRM regime, the loss landscape will be very flat within the adversarial training perturbation radius but very sharp outside the neighborhood.

Through numerical experiments, the dynamics of the measure observed in different perturbation radius δ and different training epoch shows that the behavior of models trained by adversarial training is very similar to CGRM. More details are presented in Section 5.

To further support our conjecture, we theoretically analyze the optimization process of adversarial training under a simple linear separable data set. Our results (Theorem 4.1 and Theorem 4.2) show that clean classification on large margin data implies good clean performance on small margin data, and ReLU networks with moderate size can robustly memorize small margin data efficiently. In other words, we prove that after adversarial training, the network exactly learns the function we described in Section 3.

Finally, motivated by CGRM, we propose clean and robust generalization bounds. On one hand, clean generalization bound (Theorem 6.2) shows that clean sample complexity is polynomial in data dimension d. On the other hand, we derive a upper bound of robust generalization gap (Theorem 6.3) that only relies on the number of samples and global flatness of loss landscape over input. Since adversarial training only promotes model robustly memorizing training data, the global flatness of landscape has no guarantee, and it is also verified by numerical results, which explains why robust generalization gap is large although robust training error can be very low in adversarial training,

2. Related work

Adversarial Examples and Adversarial Training (AT). Szegedy et al. (2013) first made an intriguing discovery: even state-of-the-art deep neural networks are vulnerable to adversarial examples. Then, Goodfellow et al. (2014) proposes FGSM to seek adversarial examples, and it also introduce the adversarial training framework to enhance the defence of models against adversarial perturbations. Madry et al. (2017) designs PGD to make adversarial attack and uses it to improve adversarial training, which can be viewed as the multi-step vision of FGSM. In general, FGSM-based AT and PGD-based AT are commonly regarded as the standard methods for adversarial training.

Robust Overfitting. One surprising behavior of deep learning is that over-parameterized neural networks can generalize well, which is also called *benign overfitting* that deep models have not only the powerful memorization but a good performance for unseen data (Zhang et al., 2017; Belkin et al., 2019). However, in contrast to the standard (non-robust) generalization, for the robust setting, Rice et al. (2020) empirically investigates robust performance of models based on adversarial training methods, which are designed to improve adversarial robustness (Szegedy et al., 2013; Madry et al., 2017), and the work shows that *robust overfitting* can be observed on multiple datasets.

Theoretical Understanding of Robust Overfitting. A list of works (Schmidt et al., 2018; Balaji et al., 2019; Dan et al., 2020) study the *sample complexity* for adversarial robustness, and their works manifest that adversarial robust generalization requires more data than the standard setting, which gives an explanation of the robust generalization gap from the perspective of statistical learning theory. And another line of works (Tsipras et al., 2018; Zhang et al., 2019) propose a principle called the *robustness-accuracy trade-off* and have theoretically proven the principle in different setting, which mainly explains the widely observed drop of robust test accuracy due to the trade-off between adversarial robustness and clean test accuracy. Recently, Li et al. (2022) investigates the robust expressive ability of neural networks, which demonstrates that, for the well-separated dataset, robust generalization requires exponentially large models, so the hardness of robust generalization may stem from the *expressive power* of practical models.

Memorization in Adversarial Training. Dong et al. (2021); Xu et al. (2021) empirically and theoretically explore the memorization effect in adversarial training for promoting a deeper understanding of model capacity, convergence, generalization, and especially robust overfitting of the adversarially trained models. However, different from their works, the concept *clean generalization and robust memorization (CGRM)* proposed in our paper focuses on both robust overfitting and high clean test accuracy, which means that there is surprisingly no clean memorization or clean overfitting.

3. CGRM in adversarial training

In this section, we first introduce some preliminaries in our theoretical framework.

We consider a binary classification task $\mathcal{X} \to \mathcal{Y}$, where we use $\mathcal{X} \in [0,1]^d, \mathcal{Y} = \{-1, 1\}$ to denote the supporting set of all data input and ground-truth labels, respectively, and the data input dimension is d. Let $\mathcal{D}$ be the underlying data distribution. We use clean $0-1$ loss to evaluate clean performance of classifier as

$$\mathcal{L}_{\mathcal{D}}^{\text{clean}}(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{I}\{\text{sgn}(f(x)) \neq y\}].$$

Then, we assume that data can be clean separated by a ReLU network with reasonable size (width and depth). Specifically, we assume that there exists a ReLU network classifier f_{clean} such that $\mathcal{L}_{\mathcal{D}}^{\text{clean}}(f_{\text{clean}}) = 0$, where the clean classifier is defined as

$$f_{\text{clean}}(x) := W_L\sigma(\dots\sigma(W_1x + b_1)\dots) + b_L,$$

where $W_i \in \mathbb{R}^{m_i \times m_{i-1}}, b_i \in \mathbb{R}^{m_i}, 1 \leq i \leq L, m_0 = d, m_L = 1$, and we use $\sigma(\cdot)$ to denote ReLU activation function, which is defined as $\sigma(\cdot) = \max\{0, \cdot\}$. Besides, we consider that the width $\max\{m_0, m_1, \cdots, m_L\}$ is $O(d)$, and the depth L is constant.

This consideration is taken into our setting for two reasons. First, we can use (small) networks with moderate size to achieve high test accuracy on the multiple dataset, such as MNIST and CIFAR10, so the width bounded by data dimension and the constant depth are reasonable in practice. Second, some previous theoretical work (Lu et al., 2017) shows $O(d)$ width is enough for ReLU nets to approximate a broad class of non-linear functions.

To understand adversarial robustness by a geometry way, we then introduce a notion called *decision boundary*. Formally, the decision boundary of a classifier f is defined as,

$$\begin{aligned} DB(f) := \{x \in \mathcal{X} \mid f(x) = 0, \forall \epsilon > 0, \exists x', x'' \\ \in B_p(x, \epsilon), s.t. f(x')f(x'') < 0\}. \end{aligned}$$

Notice that this definition is different from that in Zhang et al. (2019), since it requires that the sign of the neighborhood of the decision boundary can be changed, which helps us to establish the relative between decision boundary and adversarial robust margin. We define l_p adversarial margin over classifier f and data point (x, y) as $\min_{\|\xi\|_p \le \delta} yf(x + \xi)$, then we have the following proposition.

Proposition 3.1. *(The equivalent between adversarial margin and distance from decision boundary) Assume that the distance between data point and decision boundary is well-defined, then it holds that,*

$$\begin{aligned} &\{x \in \mathcal{X} \mid \min_{\|\xi\|_p \le \delta} yf(x + \xi) \ge 0\} \\ &\quad = \{x \in \mathcal{X} \mid \mathrm{dist}_p(x, DB(f)) \ge \delta\}, \end{aligned}$$

and

$$\begin{aligned} &\{x \in \mathcal{X} \mid \min_{\|\xi\|_p \le \delta} yf(x + \xi) < 0\} \\ &\quad = \{x \in \mathcal{X} \mid \mathrm{dist}_p(x, DB(f)) < \delta\}, \end{aligned}$$

where we use $\mathrm{dist}_p(\cdot, \mathcal{C})$ *to denote* l_p *distance between point* $\cdot$ *and curve* $\mathcal{C}$.

Proposition 3.1 shows that the classifier is robust on data with large distance from decision boundary, and is non-robust on data with small distance from decision boundary. Thus, according distance from decision boundary, we can divide all data into two classes, large margin data and small margin data. Namely, the former is defined as $\mathcal{X}_{\text{large}} = \{x \in \mathcal{X} \mid \mathrm{dist}_p(x, DB(f_{\text{clean}})) \ge \delta\}$, and the latter is defined as $\mathcal{X}_{\text{small}} = \{x \in \mathcal{X} \mid \mathrm{dist}_p(x, DB(f_{\text{clean}})) < \delta\}$. We also consider $\mathbb{P}_{(x,y)\sim\mathcal{D}}\{x \in \mathcal{X}_{\text{small}}\} = \mu$ and $\mathbb{P}_{(x,y)\sim\mathcal{D}}\{x \in \mathcal{X}_{\text{large}}\} = 1 - \mu$, where $0 < \mu < 1$ is a proportional constant.

Another key notion of data in our setting is well-separated, which means that data is far from each other although there exists small margin data. This property is widely observed in Yang et al. (2020), and it is clear that this assumption is foundation of robust classifier. Formally, we consider the supporting set $\mathcal{X} = A \cup B \subset [0, 1]^d$, where two disjoint sets A, B denote positive class and negative class, respectively. And we assume that $\mathrm{dist}_p(A, B) > R$, where we use $\mathrm{dist}_p(\cdot, \cdot)$ to denote l_p distance between two sets.

3.1. CGRM under the realistic data assumption

In this subsection, we consider adversarial training with access to $N-$sample training dataset

$$\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\},$$

which is i.i.d drawn from the data distribution $\mathcal{D}$, we minimize $0-1$ adversarial robust training loss as

$$\mathcal{L}_{\mathcal{S}}^{\text{adv},\delta}(f) := \frac{1}{N} \sum_{i=1}^{N} \max_{\|\xi\|_p \le \delta} \mathbb{I}\{\mathrm{sgn}(f(x_i + \xi)) \ne y_i\},$$

where δ is the perturbation radius.

Indeed, we can also divide $\mathcal{S}$ into two sets, small margin training dataset and large margin training dataset. Specifically, let $\mathcal{S}_{\text{small}}$ be the set

$$\{(x_i, y_i) \in \mathcal{S} \mid \mathrm{dist}_p(x_i, DB(f_{\text{clean}})) < \delta\},$$

and $\mathcal{S}_{\text{large}}$ be the set

$$\{(x_i, y_i) \in \mathcal{S} \mid \mathrm{dist}_p(x_i, DB(f_{\text{clean}})) \ge \delta\}.$$

To evaluate robust performance of models, we use $0-1$ adversarial robust test loss as

$$\mathcal{L}_{\mathcal{D}}^{\text{adv},\delta}(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\|\xi\|_p \le \delta} \mathbb{I}\{\mathrm{sgn}(f(x + \xi)) \ne y\} \right].$$

The final goal of adversarial training is to find good solution with low robust test loss.

Now, we present the concept that we mainly discuss in this work, *clean generalization and robust memorization (CGRM)*. Concretely, we consider the following classifier

$$\begin{aligned} f_{\text{adv}}(x) := &\sum_{(x_i, y_i) \in S_{\text{small}}} y_i \mathbb{I}\{\|x - x_i\|_p \le \delta\} \\ &+ f_{\text{clean}}(x) \mathbb{I}\left\{ \min_{(x_i, y_i) \in S_{\text{small}}} \|x - x_i\|_p > \delta \right\}. \end{aligned}$$

Under the above realistic data assumption, we can derive the following result.

Proposition 3.2. *Assume that data within the* l_p $\delta-$ *neighborhood of training data has local constant label (which means data within neighborhood of a certain data has the same label), then the classifier* f_{adv} *has the following properties:*

- *For adversarial robust training error, $\mathcal{L}_{\mathcal{S}}^{adv,\delta}(f_{adv}) = 0$;*
- *For clean test error, $\mathcal{L}_{\mathcal{D}}^{clean}(f_{adv}) = 0$;*
- *For adversarial robust test error,*

$$\mathcal{L}_{\mathcal{D}}^{adv,\delta}(f_{adv}) = \Omega\left(\mathcal{L}_{\mathcal{D}}^{adv,\delta}(f_{clean})\right) = \Omega(\mu).$$

This proposition shows that the classifier f_{adv} has the same behavior as deep models trained by adversarial training. They have the common performance that they achieve low robust training error and high clean test accuracy, but fail to robust classify the data population $\mathcal{D}$.

Motivated by this, we conjecture that adversarial training can learn functions that are close to this form, which is called *clean generalization and robust memorization (CGRM)*, an implicit bias of adversarial training. In the later sections, we formally analyze the properties of this function and adversarial training systematically from different views, including representation complexity, empirical evidence and theoretical evidence. Finally, based on CGRM, we establish generalization guarantees to explain good clean generalization and robust generalization gap.

3.2. On the representation complexity of CGRM

Next, we study the representation complexity of CGRM, which asks how many parameters are enough to express the CGRM classifier f_{adv} via deep neural networks. In fact, only $\tilde{O}(\mu N d + \text{poly}(d))$ weights are sufficient for ReLU nets to uniformly approximate the target function f_{adv}, which is stated as the following theorem.

Theorem 3.3. *Assume that the perturbation radius satisfies $\delta < R/2$, then there exists a classifier $\hat{f}$ represented by a ReLU net with at most*

$$\tilde{O}(\mu N d + \text{poly}(d))$$

parameters such that $|\hat{f} - f_{adv}| < \epsilon$ for all $x \in [0,1]^d$.

Proof sketch of Theorem 3.3 Although an exponentially large number of parameters is *necessary* to approximate a smooth function in general (DeVore et al., 1989), some simple functions can be approximate by neural networks more efficiently (Telgarsky, 2017). By leveraging this benefit, we use ReLU nets with few parameters to approximate the distance function $d_i(x) := \|x - x_i\|_p$, and we notice that the exact indicator $\mathbb{I}(\cdot)$ can be approximated by a soft indicator with two ReLU neurons. Combined with these results, we can derive Theorem 3.3.

By applying the result of Theorem 3.3, we have the following corollary, and it manifests that a neural network with $\tilde{O}(\mu N d + \text{poly}(d))$ weights can achieve CGRM.

Corollary 3.4. *Assume that the perturbation radius satisfies $\delta < R/2$, then there exists a classifier $\hat{f}$ represented by a ReLU net with $\tilde{O}(\mu N d + \text{poly}(d))$ parameters such that*

- *For robust training error, $\mathcal{L}_{\mathcal{S}}^{adv,\delta}(\hat{f}) = o(1)$;*
- *For clean test error, $\mathcal{L}_{\mathcal{D}}^{clean}(\hat{f}) = o(1)$;*
- *For robust test error,*

$$\mathcal{L}_{\mathcal{D}}^{adv,\delta}(\hat{f}) = \Omega\left(\mathcal{L}_{\mathcal{D}}^{adv,\delta}(f_{clean})\right) = \Omega(\mu).$$

However, while the previous results have shown that achieving both low robust training error and high clean test accuracy is representatively efficient for ReLU nets, robust generalization still requires exponentially large models even with our data assumption.

Theorem 3.5. *Let $\epsilon \in (0,1)$ be a small constant and $\mathcal{F}_n$ be the set of functions represented by ReLU networks with at most n parameters. There exists a sequence $N_d = \exp(\Omega(d)), d \geq 1$ and a universal constant $C_1 > 0$ such that the following holds: for any $c \in (0,1)$, there exists a underlying data distribution $\mathcal{D}$ that satisfies all above data assumptions and is μ_0-balanced, such that for any $R > 2\delta$ and robust radius $c\epsilon$, we have*

$$\inf\left\{\mathcal{L}_{D}^{adv,c\epsilon}(f) : f \in \mathcal{F}_{N_d}\right\} \geq C_1 \mu_0.$$

where μ_0-balanced means that there exists a uniform probability measure m_0 on $\mathcal{X}$ and the distribution $\mathcal{D}$ satisfies that $\inf\left\{\frac{\mathcal{D}(E)}{m_0(E)} : E \text{ is Lebesgue measurable}\right\} \geq \mu_0$.

In other words, the robust generalization error cannot be lower that a constant $\alpha = C_1 \mu_0$ unless the ReLU network has size larger than $\exp(\Omega(d))$.

The main idea of proof of Theorem 3.5 is the combination between the linear region decomposition of ReLU nets (Montufar et al., 2014) and the technique bounding the VC dimension for linear separable data that is similar to Li et al. (2022).

Therefore, we get the following representation complexity inequality,

$$\text{Representation Complexity:} \underbrace{\text{Clean Classifier}}_{\text{poly}(d)}$$
$$< \underbrace{\text{CGRM Classifier}}_{\tilde{O}(\mu N d + \text{poly}(d))} < \underbrace{\text{Robust Classifier}}_{\exp(\Omega(d))}.$$

This inequality shows that, while functions achieving CGRM have mildly higher representation complexity than clean classifiers, adversarial robustness requires excessively higher complexity, which may lead to adversarial training converges to CGRM regime.

4. Theoretical analysis of CGRM under linear-separable data

In this section, we theoretically demonstrate CGRM by analyzing the convergence of adversarial training on a certain synthetic dataset. which also indicates some critical facts that can be used to explain why the clean test performance is still good after adversarial training.

Data Distribution. We assume that data $x \in \mathbb{R}^d$ can be formalized as $cyw^* + \xi$, where w^* is the target direction ($\|w^*\|_2 = 1$), $y \in \{-1, +1\}$ is the label, c is the norm scale that is α for large margin data and is β for small margin data ($\beta \ll \alpha$), and $\xi \sim \mathcal{N}(0, \mathcal{I}_d - w^*{w^*}^{\mathrm{T}})$ is a random Guassian noise orthogonal to w^*. Indeed, this synthetic dataset captures the main characteristics of data assumption that we mention in Section 3. On one hand, there exists a simple but non-robust classifier $f_{\text{clean}}(x) = {w^*}^{\mathrm{T}}x$ that can clean classify data but is non-robust for small margin data. On the other hand, due to the randomness of noise, in the high-dimensional setting, small data is also well-separated.

Learning Model. We consider a mixed learner model as

$$f(x) = {w_0}^{\mathrm{T}}x + \sum_{i=1}^{m} a_i \phi({w_i}^{\mathrm{T}}x),$$

where the neuron $\phi(t) = \sigma(t-b)$, $\sigma(\cdot)$ is ReLU activation function and b is a threshold.

With $N-$sample training dataset $\mathcal{S} = \mathcal{S}_{\text{small}} \cup \mathcal{S}_{\text{large}}$ i.i.d. drawn from the underlying distribution $\mathcal{D}$, we minimize the exponential loss as

$$\mathcal{L}_{\mathcal{S}}^{\text{exp,adv},\delta}(f) = \frac{1}{N}\sum_{i=1}^{N} \max_{\|\delta_i\| \le \delta} \exp(-y_i f(x_i + \delta_i)).$$

By using the standard adversarial training algorithm, FGSM (Goodfellow et al., 2014), we have the following result.

Theorem 4.1. *With large margin data $\mathcal{S}_{large}$ as training data, we use FGSM to train the model f when only w_0 is activated and zero initialized, deriving a parameter iteration sequence $\{w_0^k\}_{k=1,\ldots}$. Then, with high probability over the sampled set, we have $\lim_{k\to\infty}\left\|\frac{w_0^k}{\|w_0^k\|_2} - w^*\right\|_2 = o(1)$.*

Theorem 4.1 shows that, under the linear separable data assumption, high clean test accuracy on large margin data implies good clean performance on small margin data, which can help us understand clean generalization better in adversarial training (see Theorem 6.2 in Section 6).

Theorem 4.2. *By using a modified adversarial training algorithm on all training dataset $\mathcal{S}$, we derive a parameter sequence $\{\theta^k = (w_i^k)_{i=0}^m\}_{k=1,\ldots}$. Then, with high probability, for $0-1$ adversarial training loss $\mathcal{L}_{\mathcal{S}}^{0-1,adv,\delta}$, it holds that*

$$\lim_{k\to\infty} \mathcal{L}_{\mathcal{S}}^{0-1,adv,\delta}(\theta^k) = 0.$$

In fact, the above theorems show that adversarial training method on linear separable data will converge to the normalized target solution, $f(x) = {w^*}^{\mathrm{T}}x + \sum_{x_i \in \mathcal{S}_{\text{small}}} y_i \phi(\xi_i^{\mathrm{T}}x)$, which is exactly the CGRM function mentioned in Section 3.

5. Experiments

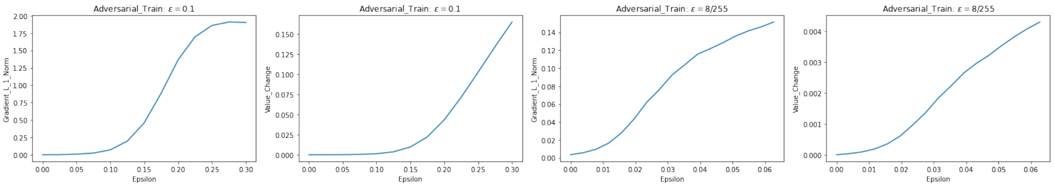
In this section, we demonstrate that, on real image datasets, adversarial training can learn classifiers in robust memorization regime. Indeed, we need to study whether models trained by adversarial training tend to memorize data points by approximating local robust indicators on training data.

Concretely, for training data (x, y), we use two measures, maximum gradient norm within the neighborhood of training data, $\max_{\|\xi\|_\infty \le \delta} \|\nabla_x \mathcal{L}(f(x+\xi), y)\|_1$ and maximum loss function value change $\max_{\|\xi\|_\infty \le \delta}[\mathcal{L}(f(x+\xi), y) - \mathcal{L}(f(x), y)]$. The former measures the $\delta-$local flatness on (x, y), and the latter measures $\delta-$local adversarial robustness on (x, y), which both describe the key information of loss landscape over input.

Experiment Settings. In numerical experiments, we mainly focus on two common real-image datasets, MNIST and CIFAR10. During adversarial training, we use cyclical learning rates and mixed precision technique (Wong et al., 2020). On MNIST, we use a LeNet5 architecture and train total 20 epochs. On CIFAR10, we use a Resnet9 architecture and train total 15 epochs.

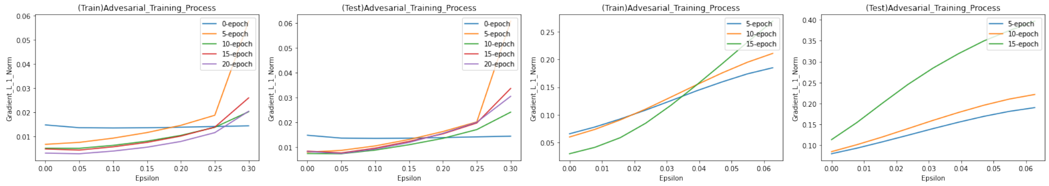
Robust Memorization. To verify robust memorization in adversarial training, we first use the standard AT to train models by a fixed perturbation radius ϵ_0, and then we compute empirical average of maximum gradient norm and maximum loss change on training data within different perturbation radius ϵ. We can see numerical results in Figure 1, and it shows that loss landscape has flatness within the training radius, but is very sharp outside, which practically demonstrates robust memorization on real image datasets, including MNIST and CIFAR10.

Learning Process. We also focus on the dynamics of loss landscape over input during the adversarial learning process. Thus, we compute empirical average of maximum gradient norm within different perturbation radius ϵ and in different training epochs. The numerical results are plotted in Figure 2. On both MNIST and CIFAR10, with epochs increasing, it is observed that the training curve descents within training perturbation radius, which implies models learn the local robust indicators to robustly memorize training data. However, while the test curve of MNIST has the similar behavior to training, on CIFAR10, the test curve ascents within training radius instead, which potentially explains why robust generalization gap on CIFAR10 is more significant than that on MNIST.



(a) MNIST: Grad (b) MNIST: Change (c) CIFAR10: Grad (d) CIFAR10: Change

Figure 1. **(a)(b):** Robust Memorization on MNIST with Training l_∞ Perturbation Radius $\epsilon = 0.1$; **(c)(d):** Robust Memorization on CIFAR10 with Training l_∞ Perturbation Radius $\epsilon = 8/255$.



(a) MNIST: Train (b) MNIST: Test (c) CIFAR10: Train (d) CIFAR10: Test

Figure 2. **(a)(b):** Learning Process on MNIST with Training l_∞ Perturbation Radius $\epsilon = 0.1$; **(c)(d):** Learning Process on CIFAR10 with Training l_∞ Perturbation Radius $\epsilon = 8/255$.

6. Generalization guarantees based on CGRM

Standard generalization bound can not directly explain high clean test accuracy after adversarial training. In general, standard generalization bound can be stated as the following form. With high probability over random sampled training data, we have

$$\mathcal{L}_{\mathcal{D}}^{\text{clean}}(f) \leq \mathcal{L}_{\mathcal{S}}^{\text{clean}}(f) + \sqrt{\frac{\text{Complexity}(F)}{N}}.$$

where N is the number of samples, and Complexity(F) denotes a complexity measure of function family F, such as VC dimension and Rademacher complexity.

However, this standard generalization bound can not explain high clean test accuracy after adversarial training. In order to have enough capacity for achieving low robust training error (i.e. $\min_{f\in F} \mathcal{L}_{\mathcal{S}}^{\text{adv},\delta}(f) = 0$), we need to set Complexity$(F) = O(Nd)$ (due to Corollary 3.4 and the relation between VC dimension and the number of parameters (Bartlett et al., 2019)), which causes the above bound too loose to use.

6.1. Improved generalization bound analysis based on CGRM

Fortunately, we notice that the CGRM function f_{adv} has much lower complexity on all large margin data $(\mathrm{poly(d)}))$ than the complexity on only sampled small large margin data ($O(Nd)$). Inspired by this, we can prove a novel generalization bound by leveraging it.

Assumption 6.1. We assume that, for any classifier f outputted by adversarial training under the realistic data assumption that we mention in Section 3, we have

$$\mathcal{L}_{\mathcal{D}_{\text{small}}}^{\text{clean}}(f) = O\left(\mathcal{L}_{\mathcal{D}_{\text{large}}}^{\text{clean}}(f)\right)$$

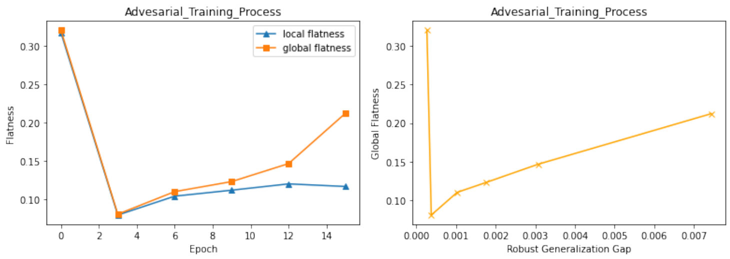
, where we use $\mathcal{D}_{\text{small}}, \mathcal{D}_{\text{large}}$ to denote small margin part and large margin part in the population $\mathcal{D}$.

This assumption has been theoretically verified in Section 4, which means that the clean test accuracy on small margin data can be bounded by the clean test accuracy on large margin data. Indeed, it holds for any homogeneous classifier when we assume that small margin data also has small norm. We leverage this property to prove the following clean generalization bound.

Theorem 6.2. *Let D be the underlying distribution that satisfies all assumption in Section 3 and Assumption 6.1. With access to $N-$sample training dataset $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ is i.i.d. drawn from $\mathcal{D}$, there exists a modified adversarial training algorithm with perturbation radius δ, $\mathcal{A}_\delta : \mathcal{D}^N \to \mathcal{F}_\mathcal{S}$, where $\mathcal{F}_\mathcal{S}$ denotes the hypothesis class, such that $\hat{f} = \mathcal{A}_\delta(\mathcal{S})$ satisfies $\mathcal{L}_{\mathcal{S}}^{adv,\delta}(\hat{f}) = 0$, and it holds with probability high probability that*

$$\mathcal{L}_{\mathcal{D}}^{clean}(\hat{f}) \lesssim \mathcal{L}_{\mathcal{S}}^{clean}(\hat{f}) + \sqrt{\frac{\text{poly}(d)}{N}}.$$

This result implies that the sample complexity of adversarial training is polynomial in the data dimension d, which provides a theoretical guarantee for benign overfitting in adversarial training.



(a) (b)

Figure 3. **Left:** Local and Global Flatness During Adversarial Training on CIFAR10; **Right:** The Relation Between Robust Generalization Gap and Global Flatness on CIFAR10.

However, we still prove the following theorems to illustrate the hardness of robust generalization. In contrast to the clean generalization, we have an another robust generalization bound that mainly depends on global flatness of loss landscape over input.

Theorem 6.3. *Let D be the underlying distribution with a smooth density function, and $N-$sample training dataset $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ is i.i.d. drawn from $\mathcal{D}$. Then, it holds with probability $1 - \Delta$ over sampled set $\mathcal{S}$ that,*

$$\mathcal{L}_{\mathcal{D}}^{adv,\delta}(f) - \mathcal{L}_{\mathcal{S}}^{adv,\delta}(f) \lesssim N^{-\frac{1}{d+2}} \left(\frac{\mathcal{C}(\mathcal{D}, \mathcal{L})}{\sqrt{\Delta}} + \underbrace{\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\|\xi\|_\infty \le \delta} \|\nabla_x \mathcal{L}(f(x+\xi), y)\|_1 \right]}_{\textit{global flatness}} \right),$$

where $\mathcal{C}(\mathcal{D}, \mathcal{L})$ is a constant that only depends on the distribution $\mathcal{D}$ and the loss function $\mathcal{L}(\cdot, \cdot)$.

This robust generalization bound shows that robust generalization gap can be controlled by global flatness of loss landscape over input rather than local flatness. And we also derive the lower bound of robust generalization gap stated as follow.

Proposition 6.4. *Let D be the underlying distribution with a smooth density function, then we have*

$$\mathcal{L}_{\mathcal{D}}^{adv,\delta}(f) - \mathcal{L}_{\mathcal{D}}^{clean}(f) = \Omega\left(\delta \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\|\nabla_x \mathcal{L}(f(x), y)\|_1 \right]\right).$$

Theorem 6.3 and Proposition 6.4 manifest that robust generalization gap is very related to global flatness. However, although adversarial training achieves good local flatness via robust memorization, models lack global flatness, which leads to robust overfitting. This point is also verified by numerical experiment on CIFAR10 (see results in Figure 3). First, global flatness grows much faster than local flatness in practice. Second, with global flatness increasing during training process, it leads to a increase of robust generalization gap.

7. Conclusion

This paper provides a theoretical understanding of adversarial training by proposing a implicit bias called robust memorization. We first explore the representation complexity of robust memorization under the realistic data assumption. Then, we empirically demonstrate robust memorization on real-image datasets. And we also theoretical analyze adversarial training in the linear separable data setting. Finally, we prove generalization guarantees inspired by robust memorization, which can explain why both good clean performance and robust overfitting happen in adversarial training.

References

Anthony, M., Bartlett, P. L., Bartlett, P. L., et al. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge, 1999.

Balaji, Y., Goldstein, T., and Hoffman, J. Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets. *arXiv preprint arXiv:1910.08051*, 2019.

Banburski, A., De La Torre, F., Pant, N., Shastri, I., and Poggio, T. Distribution of classification margins: Are all data equal? *arXiv preprint arXiv:2107.10199*, 2021.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

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

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

DeVore, R. A., Howard, R., and Micchelli, C. Optimal nonlinear approximation. *Manuscripta mathematica*, 63 (4):469–478, 1989.

Dong, Y., Xu, K., Yang, X., Pang, T., Deng, Z., Su, H., and Zhu, J. Exploring memorization in adversarial training. *arXiv preprint arXiv:2106.01606*, 2021.

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

Li, B., Jin, J., Zhong, H., Hopcroft, J. E., and Wang, L. Why robust generalization in deep learning is difficult: Perspective of expressive power. *arXiv preprint arXiv:2205.13863*, 2022.

Li, Y., Fang, E., Xu, H., and Zhao, T. Implicit bias of gradient descent based adversarial training on separable data. In *International Conference on Learning Representations*, 2020.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. The expressive power of neural networks: A view from the width. *Advances in neural information processing systems*, 30, 2017.

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

Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. On the number of linear regions of deep neural networks. *Advances in neural information processing systems*, 27, 2014.

Petzka, H., Kamp, M., Adilova, L., Boley, M., and Sminchisescu, C. Relative flatness and generalization in the interpolation regime. *arXiv preprint arXiv:2001.00939*, 2020.

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

Rice, L., Wong, E., and Kolter, Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

Silverman, B. W. *Density estimation for statistics and data analysis*. Routledge, 2018.

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

Telgarsky, M. Neural networks and rational functions. In *International Conference on Machine Learning*, pp. 3387–3393. PMLR, 2017.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

Wong, E., Rice, L., and Kolter, J. Z. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

Xu, H., Liu, X., Wang, W., Ding, W., Wu, Z., Liu, Z., Jain, A., and Tang, J. Towards the memorization effect of neural networks in adversarial training. *arXiv preprint arXiv:2106.04794*, 2021.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R. R., and Chaudhuri, K. A closer look at accuracy vs. robustness. *Advances in neural information processing systems*, 33:8588–8601, 2020.

Yarotsky, D. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

Yu, C., Han, B., Shen, L., Yu, J., Gong, C., Gong, M., and Liu, T. Understanding robust overfitting of adversarial training and beyond. In *International Conference on Machine Learning*, pp. 25595–25610. PMLR, 2022.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. iclr. *arXiv preprint arXiv:1611.03530*, 2017.

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

A. Proof of Section 3

Theorem A.1 (Restatement of Proposition 3.1)**.** *Assume that the distance between data point and decision boundary is well-defined, then it holds that,*

$$\{x \in \mathcal{X} \mid \min_{\|\xi\|_p \le \delta} yf(x+\xi) \ge 0\} = \{x \in \mathcal{X} \mid \mathrm{dist}_p(x, DB(f)) \ge \delta\},$$
$$\{x \in \mathcal{X} \mid \min_{\|\xi\|_p \le \delta} yf(x+\xi) < 0\} = \{x \in \mathcal{X} \mid \mathrm{dist}_p(x, DB(f)) < \delta\},$$

where we use $\mathrm{dist}_p(\cdot, \mathcal{C})$ *to denote* l_p *distance between point* $\cdot$ *and curve* $\mathcal{C}$.

Proof. On one hand, we assume that a data point $x \in \mathcal{X}$ with label $y = 1$ satisfies that $\mathrm{dist}_p(x, DB(f)) < \delta$. By definition of decision boundary in Section 3, we know that there exists a point $x' \in DB(f)$ such that $r = \|x - x'\|_p < \delta$, where we use r to denote the distance between x and x'. And there also exists a point x'' such that $\|x'' - x'\|_p < (\delta - r)/2$ and $f(x'') < 0$. So we have $\|x - x''\|_p < \|x - x'\|_p + \|x' - x''\|_p < \delta$ and $yf(x'') < 0$, which implies that we only need to set $\xi = x'' - x$.

On the other hand, we assume that $\min_{\|\xi\|_p \le \delta} yf(x+\xi) < 0$. Let ξ_0 be the perturbation such that $yf(x+\xi_0) < 0$. By the continuity of ReLU networks, we know that there must be a point $x' = x + \theta\xi_0$ (where $\theta \in (0,1)$) such that $x' \in DB(f)$. Thus, we complete the proof. □

Theorem A.2 (Restatement of Theorem 3.3)**.** *Assume that the perturbation radius satisfies* $\delta < R/2$*, then there exists a classifier* $\hat{f}$ *represented by a ReLU net with at most*

$$\tilde{O}(\mu N d + \mathrm{poly}(d))$$

parameters such that $|\hat{f} - f_{adv}| < \epsilon$ *for all* $x \in [0,1]^d$.

Proof. We first consider the case when $p = \infty$, and present two key lemmas to prove the theorem.

Lemma A.3 ((Yarotsky, 2017))**.** *Let* $\epsilon > 0$, $0 < a < b$ *and* $B \ge 1$ *be given. There exists a function* $\widetilde{\times} : [0,B]^2 \to [0, B^2]$ *computed by a ReLU network with* $O\left(\log^2\left(\epsilon^{-1}B\right)\right)$ *parameters such that*

$$\sup_{x,y \in [0,B]} \left|\widetilde{\times}(x,y) - xy\right| \le \epsilon,$$

and $\widetilde{\times}(x,y) = 0$ *if* $xy = 0$.

Lemma A.4. *The* l_∞ *distance function between points can be represented by ReLU nets with at most* $O(d)$ *parameters.*

Proof of Lemma A.4 Let $x^{(i)}$ denote the i-th coordinate of x, then $\|x - x_0\|_\infty = \max_{1 \le i \le d}\left|x^{(i)} - x_0^{(i)}\right|$. Since $|a| = \frac{1}{2}\left(\max\{a, 0\} + \max\{-a, 0\}\right)$, we can see that $x^{(i)} \to \left|x^{(i)} - x_0^{(i)}\right|$ can be represented by a constant-size ReLU network. Moreover, the function $\max\{a, b\} = \frac{1}{2}\left(|a+b| + |a-b|\right)$, so that the function $(a_1, a_2, \cdots, a_d) \to \max_{1 \le i \le d} a_d$ can be represented with $O(d)$ parameters. To summarize, $x \to \|x - x_0\|_\infty$ can be represented using a ReLU network of size $O(d)$, as desired.

Return to our main proof back, indeed, functions computed by ReLU networks are piecewise linear but the indicator functions are not continuous, so we need to relax the indicator such that $\hat{I}_{\text{soft}}(x) = 1$ for $x \le \delta + \epsilon_0$, $\hat{I}_{\text{soft}}(x) = 0$ for $x \ge R - \delta\epsilon_0$ and $\hat{I}_{\text{soft}}$ is linear in $(\delta + \epsilon_0, R - \delta\epsilon_0)$ by using only two ReLU neurons, where ϵ_0 is sufficient small for approximation.

Now, we notice that the robust memorization function has the following form,

$$f_{\text{adv}}(x) := \sum_{(x_i,y_i) \in S_{\text{small}}} y_i \mathbb{I}\{\|x - x_i\|_\infty \le \delta\} + f_{\text{clean}}(x)\mathbb{I}\left\{\min_{(x_i,y_i) \in S_{\text{small}}} \|x - x_i\|_\infty > \delta\right\}.$$

and f_{clean} is ReLU nets with $O(\mathrm{poly}(d)$ parameters. By applying Lemma A.3, Lemma A.4 and the above relaxed indicator, we can derive Theorem A.2.

Then, we consider the case when $p = 2$, which requires approximating l_2 distance between points instead. The following lemma solves this problem.

Lemma A.5. *The function $f(x) = x^2$ on the segment $[0, 1]$ can be approximated with any error $\epsilon > 0$ by a ReLU network having the depth and the number of weights and computation units $O(\log(1/\epsilon))$.*

Indeed, by applying the above result, we can derive the case when $p = 2$. □

Theorem A.6 (Restatement of Theorem 3.5)**.** *Let $\epsilon \in (0, 1)$ be a small constant and $\mathcal{F}_n$ be the set of functions represented by ReLU networks with at most n parameters. There exists a sequence $N_d = \exp(\Omega(d)), d \geq 1$ and a universal constant $C_1 > 0$ such that the following holds: for any $c \in (0, 1)$, there exists a underlying data distribution $\mathcal{D}$ that satisfies all above data assumptions and is μ_0-balanced, such that for any $R > 2\delta$ and robust radius $c\epsilon$, we have*

$$\inf\left\{\mathcal{L}_{D}^{adv,c\epsilon}(f) : f \in \mathcal{F}_{N_d}\right\} \geq C_1\mu_0.$$

where μ_0-balanced means that there exists a uniform probability measure m_0 on $\mathcal{X}$ and the distribution $\mathcal{D}$ satisfies that $\inf\left\{\frac{\mathcal{D}(E)}{m_0(E)} : E \text{ is Lebesgue measurable and } m_0(E) > 0\right\} \geq \mu_0$.

Proof. Here, we consider the case when $p = 2$, and the case when $p = \infty$ is similar.

First, we present two lemmas about VC dimension as follows.

Lemma A.7. *(Bartlett et al., 2019) Consider a ReLU network with L layers and W total parameters. Let F be the set of (real-valued) functions computed by this network. Then we have VC-dim$(F) = O(W \log(WL))$.*

The growth function is connected to the VC-dimension via the following lemma; see e.g. (Anthony et al., 1999).

Lemma A.8. *Suppose that VC-dim$(\mathcal{F}) = k$, then $\Pi_m(\mathcal{F}) \leq \sum_{i=0}^{k}\binom{m}{i}$. In particular, we have $\Pi_m(\mathcal{F}) \leq (em/k)^k$ for all $m > k + 1$.*

Now, we notice that ReLU networks are piece-wise linear functions. Montufar et al. (2014) study the number of local linear regions, which provides the following result.

Proposition A.9. *The maximal number of linear regions of the functions computed by any ReLU network with a total of n hidden units is bounded from above by 2^n.*

Thus, for a given clean classifier f_{clean} represented by a ReLU net with $\text{poly}(d)$ parameters, we know there exists at least a local region V such that decision boundary of f_{clean} is linear hyperplane in V. And we assume that the hyperplane is $x^{(d)} = \frac{1}{2}$.

Then, let V' be the projection of V on the decision boundary $DB(f_{\text{clean}})$, and P be an 2ϵ-packing of V'. Since the packing number $\mathcal{P}(V', \|\cdot\|, 2\epsilon) \geq \mathcal{C}(V', \|\cdot\|_2, 2\epsilon) = \exp(\Omega(d))$, where $\mathcal{C}(\Theta, \|\cdot\|, \epsilon)$ is the ϵ-covering number of a set Θ. For any $\epsilon_0 \in (0, 1)$, we can consider the construction

$$S_\phi = \left\{\left(x, \frac{1}{2} + \epsilon_0 \cdot \phi(x)\right) : x \in P\right\},$$

where $\phi : P \to \{-1, +1\}$ is an arbitrary mapping. It's easy to see that all points in S_ϕ with first $d-1$ components satisfying $\|x\|_2 \leq \sqrt{1-\epsilon_0^2}$ are in V', so that by choosing ϵ_0 sufficiently small, we can guarantee that $|S_\phi \cap V| = \exp(\Omega(d))$. For convenience we just replace S_ϕ with $S_\phi \cap V$ from now on.

Let $A_\phi = S_\phi \cap \{x \in V : x^{(d)} > \frac{1}{2}\}$, $B_\phi = S_\phi - A_\phi$. It's easy to see that for arbitrary ϕ, the construction is linear-separable and satisfies 2ϵ-separability.

Assume that for any choices of ϕ, the induced sets A_ϕ and B_ϕ can always be robustly classified with $(c\epsilon, 1-\alpha)$-accuracy by a ReLU network with at most M parameters. Then, we can construct an *enveloping network* F_θ with $M - 1$ hidden layers, M neurons per layer and at most M^3 parameters such that any network with size $\leq M$ can be embedded into this envelope network. As a result, F_θ is capable of $(c\epsilon, 1-\alpha)$-robustly classify any sets A_ϕ, B_ϕ induced by arbitrary choices of

ϕ. We use R_ϕ to denote the subset of $S_\phi = A_\phi \cup B_\phi$ satisfying $|R_\phi| = (1-\alpha)\,|S_\phi| = \exp(\Omega(d))$ such that R_ϕ can be $c\epsilon$-robustly classified.

Next, we estimate the lower and upper bounds for the cardinal number of the vector set

$$R := \{(f(x))_{x\in\mathcal{P}} | f \in F_{N_d}\}.$$

Let n denote $|\mathcal{P}|$, then we have

$$R = \{(f(x_1), f(x_2), ...f(x_n)) | f \in F_{N_d}\},$$

where $\mathcal{P} = \{x_1, x_2, ..., x_n\}$.

On one hand, we know that for any $u \in \{-1,1\}^n$, there exists a $v \in R$ such that $d_H(u,v) \leq \alpha n$, where $d_H(\cdot,\cdot)$ denotes the Hamming distance, then we have

$$|R| \geq \mathcal{N}(\{-1,1\}^n, d_H, \alpha n) \geq \frac{2^n}{\sum_{i=0}^{\alpha n}\binom{n}{i}}.$$

On the other hand, by applying Lemma A.8, we have

$$\frac{2^n}{\sum_{i=1}^{\alpha n}\binom{n}{i}} \leq |R| \leq \Pi_{F_{N_d}}(n) \leq \sum_{j=0}^{l}\binom{n}{j}.$$

where l is the VC-dimension of F_{N_d}. In fact, we can derive $l = \Omega(n)$ when α is a small constant. Assume that $l < n-1$, then we have $\sum_{j=0}^{l}\binom{n}{j} \leq (en/l)^l$ and $\sum_{i=1}^{\alpha n}\binom{n}{i} \leq (e/\alpha)^{\alpha n}$, so

$$\frac{2^n}{(e/\alpha)^{\alpha n}} \leq |R| \leq (en/l)^l.$$

We define a function $h(x)$ as $h(x) = (e/x)^x$, then we derive

$$2 \leq \left(\frac{e}{\alpha}\right)^{\alpha}\left(\frac{e}{l/n}\right)^{l/n} = h(\alpha)h(l/n).$$

When α is sufficient small, $l/n \geq C(\alpha)$ that is a constant only depending on α, which implies $l = \Omega(n)$. Finally, by using Lemma A.7 and $n = |\mathcal{P}| = \exp(\Omega(d))$, we know $N_d = \exp(\Omega(d))$.

Finally, by definition of balanced distribution, we complete the proof of Theorem A.6. □

B. Proof of Section 4

Theorem B.1 (Restatement of Theorem 4.1)**.** *With large margin data $\mathcal{S}_{\textit{large}}$ as training data, we use FGSM to train the model f when only w_0 is activated and zero initialized, deriving a parameter sequence $\{w_0^k\}_{k=1,\ldots}$. Then, with high probability over sampled set, we have*

$$\left\| \lim_{k\to\infty} \frac{w_0^k}{\|w_0^k\|_2} - w^* \right\|_2 = o(1).$$

Proof. We first introduce the following lemma, which is proposed in Li et al. (2020).

Lemma B.2. *We use N' to denote $(1-\mu)N$, and assume that $\{(x_1, y_1), (x_2, y_2), \ldots, (x_{N'}, y_{N'})\}$ is the large margin part in $\mathcal{S}$. And we defined max-margin weight as*

$$u = \operatorname*{argmax}_{\|w\|_2=1} \min_{i=1,\ldots,N'} \min_{\|\delta_i\|_2\le\delta} y_i \left(x_i + \delta_i\right)^{\mathrm{T}} w.$$

Then, we have

$$\lim_{k\to\infty} \frac{w_0^k}{\|w_0^k\|_2} = u.$$

By applying Lemma B.2, we only need to consider how to determine u in Lemma B.2.

In fact, according to Cauchy inequality, we know that we only need to solve the problem as

$$u = \operatorname*{argmax}_{\|w\|_2=1} \min_{i=1,\ldots,N'} y_i {x_i}^{\mathrm{T}} w.$$

In other words, we need to solve the max-min problem as

$$\gamma = \max_{\|w\|_2=1} \min_{i=1,\ldots,N'} y_i {x_i}^{\mathrm{T}} w.$$

Recall that $x_i = \alpha w^* + \xi_i$, where ξ_i is i.i.d. Guassian random variable. Then, we have

$$\begin{aligned}\gamma = \max_{\|w\|_2=1} \min_{i=1,\ldots,N'} y_i {x_i}^{\mathrm{T}} w &\le \max_{\|w\|_2=1} \frac{1}{N'} \sum_{i=1}^{N'} y_i {x_i}^{\mathrm{T}} w \\ &= \max_{\|w\|_2=1} \alpha {w^*}^{\mathrm{T}} w + \frac{1}{N'} \sum_{i=1}^{N'} y_i {\xi_i}^{\mathrm{T}} w.\end{aligned}$$

By using the technique of concentration, we know that, with high probability, $\|u - w^*\|_2 = o(1)$. □

Theorem B.3 (Restatement of Theorem 4.2)**.** *By using a modified adversarial training algorithm on all training dataset $\mathcal{S}$, we derive a parameter sequence $\{\theta^k = (w_i^k)_{i=0}^m\}_{k=1,\ldots}$. Then, with high probability, for $0-1$ adversarial training loss $\mathcal{L}_{\mathcal{S}}^{0-1,adv,\delta}$, it holds that*

$$\lim_{k\to\infty} \mathcal{L}_{\mathcal{S}}^{0-1,adv,\delta}(\theta^k) = 0.$$

Proof. Here, we use a modified adversarial training algorithm, which is the same as standard AT on large margin data. For small large data, the update is $w_i = \mathrm{Proj}_{{w_0}^{\perp}}(x_i)$, where x_i denotes $i-$th small margin data and $m = \mu N$.

By applying the result of Theorem 4.1, we derive that $\|w_0 - w^*\|_2 = o(1)$ and $\|w_i - \xi_i\|_2 = o(1)$. Finally, according to the properties of high dimensional probability, we complete the proof. □

C. Proof of Section 6

Theorem C.1 (Restatement of Theorem 6.2)**.** *Let D be the underlying distribution that satisfies all assumption in Section 3 and Assumption 6.1. With access to $N-$sample training dataset $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ is i.i.d. drawn from $\mathcal{D}$, there exists a modified adversarial training algorithm with perturbation radius δ, $\mathcal{A}_\delta : \mathcal{D}^N \to \mathcal{F}_\mathcal{S}$, where $\mathcal{F}_\mathcal{S}$ denotes the hypothesis class, such that $\hat{f} = \mathcal{A}_\delta(\mathcal{S})$ satisfies*

$$\mathcal{L}_{\mathcal{S}}^{adv,\delta}(\hat{f}) = 0,$$

and it holds with probability high probability over sampled set $\mathcal{S}$ that,

$$\mathcal{L}_{\mathcal{D}}^{clean}(\hat{f}) = O\left(\mathcal{L}_{\mathcal{S}}^{clean}(\hat{f}) + \sqrt{\frac{\mathrm{poly}(d)}{N}}\right).$$

Proof. The key to prove this theorem is the following modified adversarial training algorithm, which also construct a particular hypothesis class $\mathcal{F}_\mathcal{S}$ depending on specific training dataset $\mathcal{S}$.

Due to the clean classifier f_{clean} represented by ReLU net with $O(\mathrm{poly}(d))$ parameters, we know that there exists a $N_d = O(\mathrm{poly}(d))$ such that

$$\inf_{f \in \mathcal{H}_{N_d}} \{\min\{\mathcal{L}_{\mathcal{D}}^{\text{clean}}(f), \mathcal{L}_{\mathcal{D}_{\text{large}}}^{\text{adv},\delta}(f)\}\} = 0,$$

where we use $\mathcal{H}_{N_d}$ to denote the function family that can be represented by ReLU net with at most N_d parameters. It shows that $\mathcal{H}_{N_d}$ has the sufficient capacity to include the clean classifier f_{clean}.

Thus, we first use clean training on $\mathcal{S}$ with the hypothesis class $\mathcal{H}_{N_d}$ by ERM. We assume that the model learns the clean classifier $\hat{f}_{\text{clean}}$, so we can know what training data is with small margin by seeking adversarial examples on $\hat{f}_{\text{clean}}$.

Now, assume that we derive the information about $\mathcal{S}_{\text{small}}$ and $\mathcal{S}_{\text{large}}$. We design the following hypothesis class

$$\mathcal{F}_\mathcal{S} = \{\phi_1 + \tilde{\times}(\phi_2, f) \mid f \in \mathcal{H}_{N_d}\},$$

where ϕ_1 is a ReLU net for approximating $\sum_{(x_i,y_i)\in\mathcal{S}_{\text{small}}} y_i \mathbb{I}\{\|x - x_i\|_p \le \delta\}$, ϕ_2 is a ReLU net for approximating $\mathbb{I}\left\{\min_{(x_i,y_i)\in S_{\text{small}}} \|x - x_i\|_p > \delta\right\}$, and $\tilde{\times}$ is the same as that in Lemma A.3.

Indeed, $\inf_{f\in\mathcal{F}_\mathcal{S}}\{\mathcal{L}_{\mathcal{S}}^{\text{adv},\delta}(\hat{f})\} = 0$. Then, we use the standard adversarial training algorithm on $\mathcal{S}$ with the hypothesis class $\mathcal{F}_\mathcal{S}$.

Under Assumption 6.1, we have

$$\mathcal{L}_{\mathcal{D}}^{\text{clean}}(\hat{f}) = O(\mathcal{L}_{\mathcal{D}_{\text{small}}}^{\text{clean}}(\hat{f})).$$

By the conclusion of Rademacher complexity and VC dimension, with high probability over sampled set, we know

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}_{\text{small}}}^{\text{clean}}(\hat{f}) &= \mathcal{L}_{\mathcal{S}_{\text{small}}}^{\text{clean}}(\hat{f}) + O(\hat{R}_{\mathcal{S}_{\text{small}}}(\mathcal{F}_\mathcal{S})) \\
&= \mathcal{L}_{\mathcal{S}_{\text{small}}}^{\text{clean}}(\hat{f}) + O(\hat{R}_{\mathcal{S}_{\text{small}}}(\mathcal{H}_{N_d})) \\
&= \mathcal{L}_{\mathcal{S}_{\text{small}}}^{\text{clean}}(\hat{f}) + O\left(\sqrt{\frac{\mathrm{VC}_{\mathrm{d}}\mathrm{im}(\mathcal{H}_{N_d})\log(1-\mu)N}{(1-\mu)N}}\right) \\
&= \mathcal{L}_{\mathcal{S}_{\text{small}}}^{\text{clean}}(\hat{f}) + O\left(\sqrt{\frac{\mathrm{poly}(d)}{N}}\right),
\end{aligned}$$

where we use $\hat{R}_X(F)$ to denote empirical Rademacher complexity on X with F.

Thus, we complete the proof of Theorem C.1. □

Theorem C.2 (Restatement of Theorem 6.3)**.** *Let D be the underlying distribution with a smooth density function, and $N-$sample training dataset $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ is i.i.d. drawn from $\mathcal{D}$. Then, it holds with probability $1 - \Delta$ over sampled set $\mathcal{S}$ that,*

$$\mathcal{L}_{\mathcal{D}}^{adv,\delta}(f) - \mathcal{L}_{\mathcal{S}}^{adv,\delta}(f) \lesssim N^{-\frac{1}{d+2}} \left(\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\|\xi\|_\infty \le \delta} \|\nabla_x \mathcal{L}(f(x+\xi), y)\|_1\right] + \frac{\mathcal{C}(\mathcal{D},\mathcal{L})}{\sqrt{\Delta}}\right),$$

where $\mathcal{C}(\mathcal{D},\mathcal{L})$ is a constant that only depends on the distribution $\mathcal{D}$ and the loss function $\mathcal{L}(\cdot,\cdot)$.

Proof. In fact, we have the following loss decomposition,

$$\mathcal{L}_{\mathcal{D}}^{\text{adv},\delta}(f)-\mathcal{L}_{\mathcal{S}}^{\text{adv},\delta}(f)=\left(\mathcal{L}_{\mathcal{D}}^{\text{clean}}(f)-\mathcal{L}_{\mathcal{S}}^{\text{adv},\delta}(f)\right)+\left(\mathcal{L}_{\mathcal{D}}^{\text{adv},\delta}(f)-\mathcal{L}_{\mathcal{D}}^{\text{clean}}(f)\right).$$

For the first term, by using λ_i to denote kernel density estimation (KDE) in Petzka et al. (2020), then we have

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}}^{\text{clean}}(f)-\mathcal{L}_{\mathcal{S}}^{\text{adv},\delta}(f) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f(x),y)]-\frac{1}{N}\sum_{i=1}^{N}\max_{\|\xi\|\le\delta}\mathcal{L}(f(x_i+\xi,y_i)) \\
&\le \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f(x),y)]-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\xi\sim\lambda_i}[\mathcal{L}(f(x_i+\xi),y_i)] \\
&= \int_x p_{\mathcal{D}}(x)\mathcal{L}(f(x),y(x))dx-\int_x p_{\mathcal{S}}(x)\mathcal{L}(f(x),y)dx \\
&\le \underbrace{\left|\int_x (p_{\mathcal{D}}(x)-\mathbb{E}_{\mathcal{S}}[p_{\mathcal{S}}(x)])\mathcal{L}(f(x),y(x))dx\right|}_{(I)}+\underbrace{\left|\int_x (\mathbb{E}_{\mathcal{S}}[p_{\mathcal{S}}(x)]-p_{\mathcal{S}}(x))\mathcal{L}(f(x),y(x))dx\right|}_{(II)},
\end{aligned}$$

where $p_{\mathcal{S}}(x)$ is KDE of point x.

With the smoothness of density function of $\mathcal{D}$ and Silverman (2018), we know that $(I)=O(\delta^2)$.

For (II), by applying Chebychef inequality and Silverman (2018), with probability $1-\Delta$, we have

$$(II)=O(\Delta^{-\frac{1}{2}}N^{-\frac{1}{2}}\delta^{-\frac{d}{2}}+N^{-2}).$$

On the other hand, by Taylor expansion, we know

$$\mathcal{L}_{\mathcal{D}}^{\text{adv},\delta}(f)-\mathcal{L}_{\mathcal{D}}^{\text{clean}}(f)\le\delta\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\|\xi\|_\infty\le\delta}\|\nabla_x\mathcal{L}(f(x+\xi),y)\|_1\right].$$

When $\delta=N^{-\frac{1}{d+2}}$, we can derive Theorem C.2. □