The Curious Case of Adversarially Robust Models: More Data Can Help, Double Descend, or Hurt Generalization

Yifei Min[1] **Lin Chen**[2] **Amin Karbasi**[3]

[1]Dept. of Statistics and Data Science, Yale University, New Haven, CT, USA
[2]Simons Institute for the Theory of Computing, University of California, Berkeley, Berkeley, CA, USA
[3] Dept. of Electrical Engineering, Computer Science, Statistics and Data Science, Yale University, New Haven, CT, USA

Abstract

Adversarial training has shown its ability in producing models that are robust to perturbations on the input data, but usually at the expense of a decrease in the standard accuracy. To mitigate this issue, it is commonly believed that more training data will eventually help such adversarially robust models generalize better on the benign/unperturbed test data. In this paper, however, we challenge this conventional belief and show that more training data can hurt the generalization of adversarially robust models in classification problems. We first investigate the Gaussian mixture classification with a linear loss and identify three regimes based on the strength of the adversary. In the weak adversary regime, more data improves the generalization of adversarially robust models. In the medium adversary regime, with more training data, the generalization loss exhibits a double descent curve, which implies the existence of an intermediate stage where more training data hurts the generalization. In the strong adversary regime, more data almost immediately causes the generalization error to increase. Then we analyze a two-dimensional classification problem with a 0-1 loss. We prove that more data always hurts generalization of adversarially trained models with large perturbations. Empirical studies confirm our theoretical results.

1 INTRODUCTION

In recent years, modern machine learning methods have exhibited their superiority over traditional models in an abundance of machine learning tasks, e.g., image classification [Krizhevsky et al., 2012], speech recognition and language translation [Graves et al., 2013, Bahdanau et al., 2015], medical diagnosis [Lakhani and Sundaram, 2017, Xiao et al., 2019], text recognition and information extraction [Long et al., 2020, Mei et al., 2018, Wang et al., 2012], online fraud detection [Pumsirirat and Yan, 2018], and self-driving cars [Ramos et al., 2017], among others. However, they can also be extremely vulnerable to adversarial, human-imperceptible data modifications [Szegedy et al., 2014, Carlini and Wagner, 2018, Kos et al., 2018]. This vulnerability is even more concerning and dangerous when machine learning methods are used in scenarios directly connected to human safety such as medical diagnosis (misinterpreting medical images) or self-driving cars (misreading traffic signs). To circumvent these issues, practitioners introduce adversarial training in order to produce adversarially robust models [Huang et al., 2015, Shaham et al., 2018, Madry et al., 2018, Zhang et al., 2019a, Gao et al., 2019, Song et al., 2019] that can still make consistently correct predictions, even when faced with perturbed data.

There is a large body of work dedicated to adversarially robust models [Zhang and Zhu, 2019, Santurkar et al., 2019, Zhang et al., 2019b, Diochnos et al., 2019, Wei and Ma, 2019, Zhai et al., 2019]. In particular, it has been shown that there exists a trade-off between the generalization of a model (i.e. the standard accuracy) and its robustness to adversarial perturbation [Tsipras et al., 2019]. Along a similar vein, Schmidt et al. [2018] showed that adversarially robust models need more training data compared to their standard counterparts in order to achieve the same generalization performance. In this paper, we want to further investigate these ideas and explore whether simply adding more data is enough for adversarially robust models to catch up to the generalization ability of their standard counterparts.

Previous works have studied the generalization of adversarially robust models from a variety of perspectives. For instance, Yin et al. [2019], Khim and Loh [2018] and Awasthi et al. [2020] gave bounds on the generalization error of adversarially robust models via Rademacher complexity. Bhagoji et al. [2019] and Pydi and Jog [2020] studied the problem from the view of optimal transport. More recently, Chen et al. [2020b] studied the influence of a larger training

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

set upon the gap between the generalization performance of an adversarially robust model and a standard model. They proved that more training data could result in expansion of the gap and denied the belief that more training data always helps adversarially robust models reach a similar generalization performance to the standard model. Building on these works, our goal is to move past bounds and gaps, and directly characterize how the size of training set affects the accuracy of adversarially robust models on unperturbed test data.

1.1 OUR CONTRIBUTIONS

A conventional wisdom in machine learning is that a larger training set will result in better generalization on the test data. We provably establish a surprising, and to some extent even paradoxical, result that more training data can hurt the generalization of adversarially robust models. We first consider a linear classification problem with a linear loss function and identify three regimes of different adversary strengths, i.e., the weak, medium, and strong adversary regimes. The strength of the adversary here refers to the magnitude of the perturbation allowed.

- In the **strong adversary regime**, the generalization of adversarially robust models deteriorates with more training data, except for a possible short initial stage where the generalization is improved with more data.
- The **medium adversary regime** is probably the most interesting one among the three regimes. In this regime, the evolution of the generalization performance of adversarially robust models could be a double descent curve. In particular, at the initial stage, the generalization loss on the test data is reduced with more training data. At the intermediate stage, however, the generalization loss increases as there is more training data (more data hurts the generalization of adversarial robust models). At the final stage, more training data improves the generalization performance.
- In the **weak adversary regime**, the generalization is consistently improved with more training data.

We then move to the analysis of the 0-1 loss and investigate a two-dimensional classification problem where the candidate decision boundary is given by a piecewise constant function. Similar weak and strong adversary regimes are observed under this setting. In particular, in the strong adversary regime, more data always hurts the generalization of adversarially robust models.

We complement the above theroetical results with empirical studies on important machine learning models, including support vector machines (SVMs), linear regression, and Gaussian mixture classification with 0-1 loss. We observe a similar phenomenon that more data hurts generalization in adversarial training. These empirical results suggest that the observed phenomenon may be ubiquitous across different models and loss functions and that we need to reflect on the true role that the size of the training set plays in adversarial training.

2 RELATED WORK

In this section, we briefly discuss some additional papers on the generalization of adversarially robust models and the double descent phenomenon, which are most relevant to our work.

Schmidt et al. [2018] showed that adversarially robust models need more training data compared to their standard counterpart. They considered a Gaussian mixture model similar to ours and proved that the training of a robust model requires a training set with size $\Omega(d)$ where d is the dimension of the data, whereas the standard model only needs a constant number of data points. Xie et al. [2020] practically showed that adversarial examples sometimes can help standard generalization under certain cases, and pointed out that this requires the adversarial examples to be used in a right manner. Their findings indicate a complicated connection between the standard accuracy and adversarial training. Bubeck et al. [2019] studied a binary classification problem under a statistical query setting and showed that to train a robust classifier one needs exponentially (in dimension d) many queries, while only polynomially many to train a standard classifier. The main difference between their work and our work is that we quantify the training dynamic in terms of the size of the training set. Very recently, Javanmard et al. [2020] precisely characterized the trade-off of standard/robust accuracy under the linear regression setting. Raghunathan et al. [2019] gave empirical evidence that adversarial training could hurt the standard accuracy, despite its improvement on robustness. The PAC-learning setting has also been studied by several authors [Cullina et al., 2018, Diochnos et al., 2019, Montasser et al., 2020]. Cullina et al. [2018] provided a polynomial (in the VC dimension) upper bound for the sample complexity, while Diochnos et al. [2019] gave a lower bound for the sample complexity which is exponential in the dimension of the input.

The strength of the adversary is crucial in the adversarial training. Theoretically, Dohmatob [2019] showed that a classifier with high standard accuracy can inevitably be fooled by a strong adversary. Empirically, Papernot et al. [2016] and Tsipras et al. [2019] found that a strong adversary can drive down standard accuracy for robust models. Ilyas et al. [2019] found that the adversarial training tends to learn non-robust features and omit robust ones if the adversary is too strong. There has also been some work in mitigating the reduction in the standard accuracy. Empirically, it was shown that if the perturbation is relatively small and does not push the data across the decision boundary, then the generaliza-

tion can be improved [Stutz et al., 2019]. Moreover, using specially chosen adversarial examples to do the adversarial training can also be helpful [Zhang et al., 2020].

The double/multiple descent phenomenon has been studied by several authors. Belkin et al. [2019a,b], Mei and Montanari [2019], Chen et al. [2020a] provably showed the existence of double/multiple descent curves for the generalization error. However, we would like to remark that the double/multiple descent curve that they considered is in terms of the number of parameters (model complexity), while ours is sample-wise. Empirically, Nakkiran et al. [2019] also discovered a sample-wise double descent phenomenon.

In a concurrent and independent work, Raghunathan et al. [2020] performed a finite-sample analysis of the trade-off between the robustness and the standard accuracy. They considered using the robust self-training estimator [Carmon et al., 2019, Najafi et al., 2019] to mitigate the robust error without sacrificing the standard accuracy. As a comparison, they studied a linear regression model while our focus is on classification problems. In their setting, the original training dataset was augmented with perturbed examples and they investigated a regime where the optimal predictor has zero standard and robust error. Our analysis covers the magnitude of the perturbation changing from small (i.e. the weak regime) to large (i.e. the strong regime), demonstrating the standard test performance of robust classifiers trained under different regimes.

3 PRELIMINARIES

Throughout this paper, let $[n]$ be a shorthand notation for $\{1, 2, \ldots, n\}$. Assume the data point (x, y) consists of the input variable x and label y, and (x, y) is generated from some distribution $\mathcal{D}$. Denote the loss function by $\ell(x, y; w)$ and the robust classifier is defined as follows Goodfellow et al. [2015], Madry et al. [2018]:

$$w_n^{\text{rob}} = \arg\min_{w\in\Theta} \sum_{i=1}^{n} \max_{\tilde{x}_i \in B_{x_i}^{\infty}(\varepsilon)} \ell(\tilde{x}_i, y_i; w)\,, \qquad (1)$$

where Θ is the parameter space and $B_x^{\infty}(\varepsilon) := \{\tilde{x} \in \mathbb{R}^d | \|\tilde{x} - x\|_{\infty} \le \varepsilon\}$ is an ℓ^{∞} ball centered at x with radius ε. The radius ε characterizes the strength of the adversary. A larger ε means a stronger adversary. This robust classifier minimizes the robust loss, or equivalently, maximizes the robust reward (i.e., negative loss).

The generalization error of the robust classifier is given by

$$L_n = \mathbb{E}_{\{(x_i,y_i)\}_{i=1}^n} \left[\mathbb{E}_{(x,y)}[\ell(x, y; w_n^{\text{rob}})]\right]\,, \qquad (2)$$

where the inner expectation is over the randomness of the test data point $(x, y) \sim \mathcal{D}_{\mathcal{N}}$ and the outer expectation is over the randomness of the training dataset $\{(x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}_{\mathcal{N}}$. The test and training data are assumed to be independently sampled from the same distribution. The generalization error can be interpreted as the expected loss of the robust model over standard/unperturbed test data.

4 THEORETICAL RESULTS

In this section we study two different binary classification models. In Section 4.1, we analyze the Gaussian mixture model under linear loss and prove the existence of three possible regimes (weak, medium and strong adversary regimes), in which more training data can help, double descend, or hurt generalization of the adversarially trained model, respectively. In Section 4.2, we construct a model called the Manhattan model that enables us to analyze the 0-1 loss and prove that analogous weak and strong adversary regimes also exist under a different loss function.

4.1 GAUSSIAN MIXTURE WITH LINEAR LOSS

In this subsection, we consider the Gaussian mixture model with linear loss. More specifically, the distribution for the data generation is specified by $y \sim \text{Unif}(\{\pm 1\})$ and $x \mid y \sim \mathcal{N}(y\mu, \Sigma)$, where $\mu(j) \ge 0$ for all $j \in [d]$ and $\Sigma = \text{diag}(\sigma^2(1), \sigma^2(2), \ldots, \sigma^2(d))$. In the remaining parts we denote this distribution by $(x, y) \sim \mathcal{D}_{\mathcal{N}}$. We consider the linear loss $\ell(x, y; w) = -y\langle w, x\rangle$ and we set the constraint set as $w \in \Theta = \{w \in \mathbb{R}^d | \|w\|_{\infty} \le W\}$ where W is a positive constant, similar to [Chen et al., 2020b, Yin et al., 2019, Khim and Loh, 2018]. In this setting, by (1) the robust classifier is

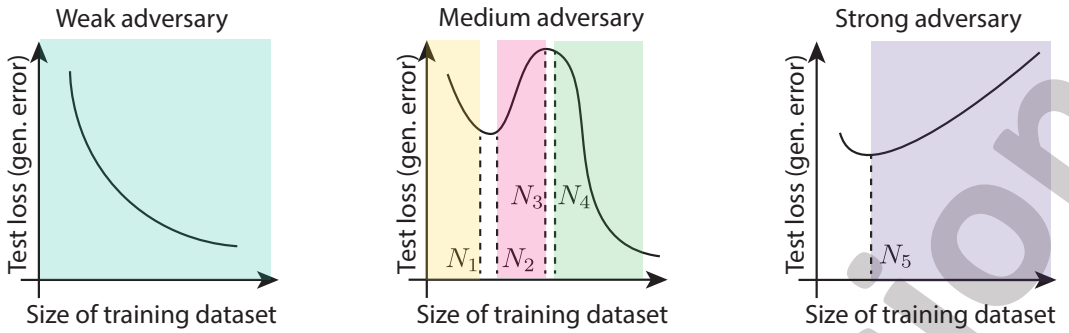
$$\begin{aligned} w_n^{\text{rob}} &= \arg\min_{\|w\|_{\infty}\le W} \sum_{i=1}^{n} \max_{\tilde{x}_i \in B_{x_i}^{\infty}(\varepsilon)} (-y_i\langle w, \tilde{x}_i\rangle) \\ &= \arg\max_{\|w\|_{\infty}\le W} \sum_{i=1}^{n} \min_{\tilde{x}_i \in B_{x_i}^{\infty}(\varepsilon)} y_i\langle w, \tilde{x}_i\rangle\,. \end{aligned} \qquad (3)$$

We study how the generalization error of the robust model evolves as the size of the training dataset changes, i.e., the dependence of L_n on n. By (2) the generalization error of the robust classifier under linear loss is given by

$$L_n = \mathbb{E}_{\{(x_i,y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}_{\mathcal{N}}} \left[\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathcal{N}}}[-y\langle w_n^{\text{rob}}, x\rangle]\right]\,. \qquad (4)$$

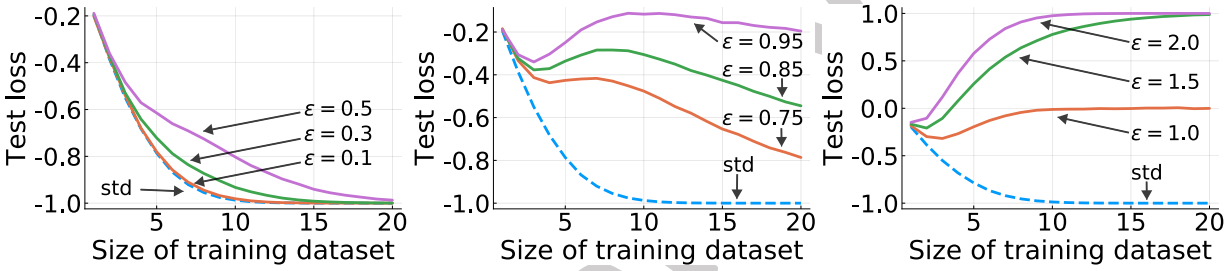
For the Gaussian classification problem under the linear loss, we identify that the behavior of L_n exhibits a phase transition which is determined by the strength of the adversary. Our main result is summarized by Theorem 1.

Theorem 1 (Proof in Section 1 of the supplementary material)**.** *Given n i.i.d. training data points $(x_i, y_i) \sim \mathcal{D}_{\mathcal{N}}$, if the robust classifier is defined by* (3) *and its generalization error is defined by* (4)*, then there exist $0 < \delta_1 < \delta_2 < 1$, such that*



(a) Weak adversary (b) Medium adversary (c) Strong adversary

Figure 1: This cartoon illustrates the three adversary regimes (i.e., weak, medium, and strong) and the corresponding results of Theorem 1. In the weak adversary regime, more training data always improves generalization. The medium adversary regime exhibits a double descent curve. When the size of training set $n \le N_1$ (the initial stage), more training data improves the generalization; when $N_2 < n < N_3$ (the intermediate stage), generalization is hurt by more data; when $n \ge N_4$, more data helps with generalization again. In the strong regime, generalization deteriorates with more data when the size of training size is sufficiently large.



(a) Weak adversary (b) Medium adversary (c) Strong adversary

Figure 2: The test loss versus the size of the training dataset under the linear loss and the one-dimensional ($d = 1$) Gaussian data generation model described in Section 4.1. The parameters of the Gaussian data model are set as follows: $\mu_0 = 1$ and $\sigma_0 = 2$. In each plot, the solid curves correspond to robust models and the dashed curve corresponds to the standard model.

(a) If $0 < \varepsilon < \delta_1 \cdot \min_{j\in[d]} \mu(j)$*, then* $L_n < L_{n-1}$ *for all* n*. That is, the loss* L_n *monotonically decreases as the number of training points* n *increases.*

(b) If $\delta_2 \cdot \max_{j\in[d]} \mu(j) < \varepsilon < \min_{j\in[d]} \mu(j)$*, and we further assume that* $\frac{\mu(j)}{\sigma(j)}$ *is the same for all* j*, then there exist* $N_1 < N_2 < N_3 < N_4$ *such that*

$$L_n \begin{cases} < L_{n-1} & \text{for } 0 < n \le N_1\,, \\ > L_{n-1} & \text{for } N_2 < n < N_3\,, \\ < L_{n-1} & \text{for } N_4 \le n\,. \end{cases}$$

(c) If $\max_{j\in[d]} \mu(j) \le \varepsilon$*, then there exists* N_5 *such that* $L_n > L_{n-1}$ *for all* $n > N_5$*.*

Remark 1. In part (b), we assume that $\mu(j)/\sigma(j)$ is the same for all j. This assumption is only for the convenience of the proof. In general, the medium regime (i.e., double descent) still exists if the assumption is relaxed. Indeed, Lemma 3 in the appendix shows that the generalization curve L_n can be decomposed as a weighted sum of losses $L(v_j, \epsilon'_j)$ over all dimensions $j \in [d]$. Therefore, if we have different ratios $\mu(j)/\sigma(j)$ for each j, then the curves $L(v_j, \epsilon'_j)$ will have different thresholds N_2 and N_3 for each j. When the d curves are added together, there will be an increasing stage in the intersection of the increasing stages of the majority of curves.

Theorem 1 proves the existence of three possible regimes during the commonly used adversarial training procedure and gives conditions for when the phase transition between these regimes will take place. Part (a) identifies the weak regime, showing that when the strength of the adversary ε is small compared to the signal μ, the generalization error decreases as the size of the training dataset increases. In this regime, the generalization benefits from the use of a large training set. This regime is illustrated by Fig. 1a, where the curve is always decreasing.

However, as the adversary becomes stronger, we reach the medium regime and things change. Part (b) proves the existence of a double descent curve for the generalization error. It shows that when ε becomes larger and approaches the signal in magnitude, the generalization error will first de-

crease as more training data is used. Surprisingly, once it reaches a certain point, it will start increasing as we feed more data. This increasing stage continues until the dataset size reaches some threshold N_2 and then the error will decrease again. The medium adversary regime is illustrated by Fig. 1b, where the three stages are marked by three different colored areas. We would like to provide high-level intuition for this regime. On one hand, a larger training dataset provides the adversary with more data to corrupt (which is bad). On the other hand, with more data, the empirical risk approximates the population risk better (which is good). Small perturbation magnifies its positive influence while large perturbation magnifies the negative influence. Medium perturbation makes both influences comparable. The medium regime happens when the negative influence prevails over the positive one at a temporary intermediate stage.

If the adversary's strength reaches the signal level or becomes even stronger, then for all sufficiently large n, the generalization error monotonically increases as the size of training set increases. This strong regime is described in part (c) of Theorem 1 and illustrated by Fig. 1c. Note that despite the decreasing stage near the very beginning, the loss keeps going up after the threshold N_5.

Based on our findings, we believe the signal-to-perturbation ratio is the key to the non-monotonicity. We hypothesize that there is a trade-off between the data size and the adversary's power to perturb the data. Specifically, given more data, the model tends to learn better. However, this also means the adversary can have more data to manipulate with. Therefore, when the ratio is large, the learner wins. When the ratio is low, the adversary has the advantage.

We would like to remark that the part for the strong regime is added mainly for the purpose of completeness of the results. It should be mentioned that the magnitude of adversarial perturbation usually does not exceed the signal level in practice. We also remark that although our theoretical results prove such non-monotonicity of the standard error versus the sample size in the medium and strong regime (i.e., when the perturbation is no longer negligible compared to the signal), in practice, similar phenomena have been observed in very realistic settings such as MNIST under quite appropriate perturbation level (see, for example, figure 1(a) in [Tsipras et al., 2019] and figure 1 in [Raghunathan et al., 2020]).

Furthermore, we see that in the medium regime, the length of the increasing stage is given by $N_3 - N_2$, according to part (b) of Theorem 1. We would like to remark that the model can have an arbitrarily long increasing stage, which depends on the adversary's strength. To better interpret this idea and the meaning behind Theorem 1, we consider the following special case where $\mu(j) = \mu_0$ and $\sigma(j) = \sigma_0$ for all $j \in [d]$. In this special case, it can be shown that in the medium regime, as ε approaches the signal strength μ_0, the increasing stage grows and can be arbitrarily long.

Corollary 2 (Proof in Section 1)**.** *Under the same assumption as Theorem 1 and further assuming that $\mu(j) = \mu_0$ and $\sigma(j) = \sigma_0$ for all $j \in [d]$, we have*

(a) If $0 < \varepsilon < \delta_1\mu_0$, then $L_n < L_{n-1}$ for all n.

(b) If $\delta_2\mu_0 < \varepsilon < \mu_0$, then there exist $N_1(\varepsilon) < N_2(\varepsilon)$ such that

$$L_n \begin{cases} < L_{n-1} & \text{for } 0 < n \le N_1\,, \\ > L_{n-1} & \text{for } N_1 < n < N_2\,, \\ < L_{n-1} & \text{for } N_2 \le n\,, \end{cases}$$

and $\lim_{\varepsilon \to \mu_0^-} N_2(\varepsilon) - N_1(\varepsilon) = +\infty$.

(c) If $\mu_0 \le \varepsilon$, then there exists $N_3(\varepsilon)$ such that $L_n > L_{n-1}$ for all $n > N_3$.

Part (a) and (c) of Corollary 2 are a re-statement of corresponding parts of Theorem 1 in the simplified setting. Part (b) additionally states that as ε increases towards μ_0, the length of the increasing stage goes to infinity. In this setting, the three regimes are marked by the thresholds $\delta_1\mu_0$, $\delta_2\mu_0$ and μ_0.

Fig. 2 illustrates the behavior of the generalization error in this simplified setting. In the simulation we set the parameters as $d = 1$, $\mu_0 = 1$ and $\sigma_0 = 2$ (for all three plots). Fig. 2a shows the weak adversary regime. We see that the generalization error maintains a decreasing trend when ε is as large as half the signal strength. In Fig. 2b, it is clear that the generalization error has a double descent curve. At first there is a decreasing stage, which is followed by an increasing stage. Also observe that as ε becomes larger, the error increases faster during the increasing stage. The error will finally start decreasing as the size of training dataset reaches the second decreasing stage. On the contrary, in the strong adversary regime, the increasing stage lasts forever and the error keeps increasing no matter how much data is provided, as illustrated by Fig. 2c.

4.2 MANHATTAN MODEL

In general, the 0-1 loss is mathematically intractable for most data models and computationally prohibitive to optimize in practice. With this in mind, we introduce a conceptual classification model that we call the Manhattan model. Note that this model is highly simplified and thus unlikely to be suitable for modeling real-world problems. Instead, the purpose of the Manhattan model is to allow a mathematical study of the 0-1 loss, and thus provide a springboard for the study of 0-1 loss in more complicated models.

We start by describing the data distribution. Assume we have data points $(x, y) \in \mathbb{R}^2 \times \{\pm 1\}$, where the support of x is given as $x = (s, t) \in \{(i, y\mu) \,:\, i \in [N],\ y \in \{\pm 1\}\}$, where $0 < \mu < 1/4$. In other words, every data point (x, y)

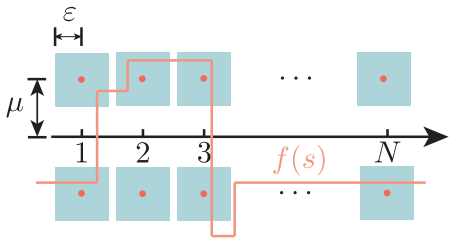


Figure 3: An illustration of the Manhattan model. The data is in the $\mathbb{R}^2$ plane and the support is $2N$ points on opposite sides of the axis. The distance between any point and the axis is μ, and the shaded square denotes the ε perturbation. The red curve shows a possible classifier.

consists of a positive or negative label y and a point on the 2-D plane $x = (s, t)$ where s is an integer between 1 and N and t is either μ or $-\mu$ depending on whether the label y is $+1$ or -1. Thus, the support consists of exactly $2N$ points with half in the positive class and half in the negative class. The data is uniformly sampled from these $2N$ points and this distribution is denoted by $\mathcal{D}_{2N}$.

Next, we consider a conceptual classifier of the form of a step function over the 2-D (s, t)-plane. That is, a classifier is defined by a function $t = f(s)$ such that $f \in F$ where

$$F = \{f : f(s) = \sum_{j=1}^{M} \alpha_j \mathbb{1}[s \in I_j],\ M \in \mathbb{N},\ \alpha_j \in \mathbb{R},\ I_j \subseteq \mathbb{R} \text{ are intervals}\}.$$

A point $x = (s, t)$ is classified $+1$ if $t > f(s)$ and -1 if $t < f(s)$. If $t = f(s)$, then x is classified as either $+1$ or -1 uniformly at random. Fig. 3 illustrates the support of the data distribution, as well as a possible classifier $f(s)$.

Since we consider the 0-1 loss, one can note that for this data model, there can be infinitely many classifiers. For example, $f(s) = c$ can attain 100% standard accuracy for all $c \in (-\mu, \mu)$. Therefore, we add an infinitesimal ℓ_1 penalty to the 0-1 loss for the purpose of tie-break, i.e., making the minimizer unique. This penalized 0-1 loss of a classifier $f(\cdot)$ can then be written as $H\left(-y(t - f(s))\right) + \lambda\|f\|_1$ with $\lambda \to 0$. For a given training set $\{(x_i, y_i) : i \in [n]\}$, We define the robust classifier over this training set as

$$f_n^{\text{rob}} \in \lim_{\lambda \to 0^+} S(\lambda)\,, \tag{5}$$

where $S(\lambda)$ is defined by

$$\arg\min_{f \in F} \sum_{i=1}^{n} \max_{\|\tilde{x}_i - x_i\|_\infty < \varepsilon} H\left(-y_i(\tilde{t}_i - f(\tilde{s}_i))\right) + \lambda\|f\|_1,$$

and

$$H(s) = \mathbb{1}[s > 0] + \frac{1}{2}\mathbb{1}[s = 0]$$

is the Heaviside step function.

Note that the RHS of Eq. (5) is the limit of a sequence of sets. This slight abuse of notation is justified by the following Lemma 3, which shows for all sufficiently small λ, the set $S(\lambda)$ remains fixed. We define the set of candidate classifiers without the penalty as

$$S = \arg\min_{f \in F} \sum_{i=1}^{n} \max_{\|\tilde{x}_i - x_i\|_\infty < \varepsilon} H\left(-y_i(\tilde{t}_i - f(\tilde{s}_i))\right),$$

and we have the following lemma.

Lemma 3 (Proof in Section 2 of the supplementary material)**.** *For all sufficiently small $\lambda > 0$ and for any $\varepsilon < 1/2$, the set $S(\lambda)$ defined by Eq.* (5) *is equivalent to the following set which is nonempty*

$$S_2 = S \cap \arg\min_{f \in S} \|f\|_1. \tag{6}$$

Lemma 3 shows that by picking a small enough λ, the minimizers with ℓ_1 penalty actually coincide with the minimizers under 0-1 loss with the smallest ℓ_1 norm. Therefore, $\lim_{\lambda \to 0^+} S(\lambda) = S_2$ and we can write $f^{\text{rob}} \in S_2$ as an equivalent definition of the robust classifier to Eq. (5).

The generalization error of f_n^{rob} is then given by

$$L_n = \mathbb{E}_{\{(x_i, y_i)\}_{i=1}^n}\left[\mathbb{E}_{(x,y)} H\left(-y\left(t - f_n^{\text{rob}}(s)\right)\right)\right], \tag{7}$$

where the inner expectation is taken over the test data point $(x, y) \sim \mathcal{D}_{2N}$, and the outer expectation is taken over the training data $\{(x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}_{2N}$. Note that here we can get rid of the ℓ_1 term due to Lemma 3. Theorem 4 shows that the generalization error can be zero for all n when $\varepsilon < 2\mu$ and can increase with n as $\varepsilon > 2\mu$.

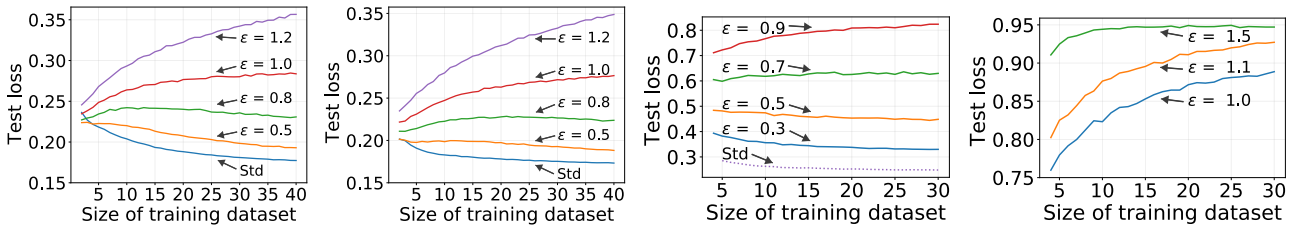
Theorem 4 (Proof in Section 3 of the supplementary material)**.** *Assume the training data $(x_i, y_i) \sim \mathcal{D}_{2N}$ where $i \in [n]$. For the robust classifier defined by* (5) *and its generalization error defined by* (7)*, we have*

(a) If $0 < \varepsilon < 2\mu$, then $L_n = 0$ for all n.
(b) If $2\mu < \varepsilon \leq 1/2$, then $L_{n+1} > L_n$ for all $n \geq 1$.

Again, the purpose of the Manhattan model is not to model any real-world problems, but instead to show that adversarial training under a 0-1 loss can also be characterized with weak/strong regimes. More generally, we have now shown that the existence of weak/strong regimes is not solely an artifact of the linear loss used in Section 4.1, and thus that it may not be surprising to see analogous results for a much broader class of loss functions.

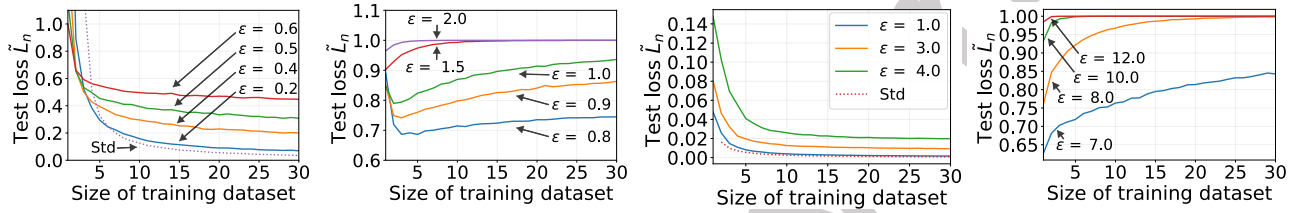
5 EMPIRICAL RESULTS

In this section, we empirically study the generalization error of robust models in three settings.



(a) Agnostic tiebreak (b) Optimal tiebreak (c) SVM (small ε) (d) SVM (large ε)

Figure 4: Figs. 4a and 4b present the test loss vs. the size of the training dataset for Gaussian mixture in the 0-1 loss setting described in Section 5.1. Figs. 4c and 4d illustrate the test loss vs. the size of the training dataset for the support vector machine model described in Section 5.2.



(a) Gaussian, small ε (b) Gaussian, large ε (c) Poisson, small ε (d) Poisson, large ε

Figure 5: The test loss versus the size of the training dataset under 1-dim linear regression model with the squared loss. The data generation follows either a Gaussian or Poisson distribution: in Fig. 5a and Fig. 5b, $x \sim \mathcal{N}(0,1)$; in Fig. 5c and Fig. 5d, $x \sim \text{Poisson}(5) + 1$. The solid curves correspond to robust models and the dashed curve corresponds to the standard model. Here $\tilde{L}_n = (L_n - \mathbb{E}e^2)/\mathbb{E}x^2$ is the scaled test loss.

5.1 GAUSSIAN MIXTURE WITH 0-1 LOSS

We consider the 1-dimensional Gaussian mixture model with 0-1 loss. The data generation is the same as in Section 4.1 with $d = 1$. Here we set $\Theta = \mathbb{R}$ and the classifier is represented by a real number $w \in \mathbb{R}$. That is, a point is classified as positive or negative depending on whether x is greater than or less than w. If $x = w$, it is uniformly randomly classified as positive or negative. Given a data point (x, y), the 0-1 loss of classifier w is given by $\ell(x, y; w) = \mathbb{1}[y(x - w) < 0]$.

We remark that under this setting, the robust classifier is not unique and the set of classifiers is an interval. As a consequence, in order to select a classifier, we need to use some tiebreaking methods. To see this, let the training dataset be $\{(x_i, y_i)\}_{i=1}^n$ and we define the neuralized dataset $\{(x_i', y_i)\}_{i=1}^n$ that satisfies $x_i' = x_i - y_i\varepsilon$ for all $i \in [n]$. In other words, for a positive sample $(x_i, y_i = 1)$, we obtain its neutralized sample by shifting x_i to the negative direction by ε, i.e., $x_i' = x_i - \varepsilon$; for a negative sample $(x_i, y_i = -1)$, its neutralized sample is obtained by shifting x_i to the positive direction by ε, i.e., $x_i' = x_i + \varepsilon$. With this definition, the robust classifier can be expressed as the following.

Proposition 5 (Proof in Section 4.1 of the supplementary material)**.** *Given the training dataset $\{(x_i, y_i)\}_{i=1}^n$ and the neuralized dataset $\{(x_i', y_i)\}_{i=1}^n$, the robust classifier is given by*

$$w_n^{\text{rob}} \in \arg\min_{w \in \mathbb{R}} \sum_{i=1}^n y_i \mathbb{1}[x_i' < w]\,. \tag{8}$$

Now one can see the tiebreaking issue in light of Proposition 5. To see this, let s be the permutation of $[n]$ such that $x'_{s(1)} \le x'_{s(2)} \le \cdots \le x'_{s(n)}$. The n points divide the real line into $n+1$ intervals: $(-\infty, x'_{s(1)}]$, $(x'_{s(i)}, x'_{s(i+1)}]$ for $1 \le i \le n-1$, and $(x'_{s(n)}, \infty)$. Let w^* be a minimizer of (8). If w^* lies in any of the above $n+1$ intervals, then any other point in the same interval is also a minimizer, since at these two points the objective function has the same value. Therefore, a tiebreaking procedure is required here.

Thus to select a classifier, we consider two tiebreaking methods. One is the agnostic tiebreak, which means the classifier is chosen uniformly at random from the interval. The other is the optimal tiebreak in hindsight, referring to picking the classifier from the interval with the smallest expected test loss.

For the agnostic tiebreak, if $w^* \in (x'_{s(i)}, x'_{s(i+1)}]$, it chooses w_n^{rob} uniformly at random from the interval. If $w^* > x'_{s(n)}$, it chooses w_n^{rob} arbitrarily close to $x'_{s(n)}$ from above. If $w^* \le x'_{s(1)}$, it chooses $w_n^{\text{rob}} = x'_{s(1)}$.

For the optimal tiebreak in hindsight, we first note that the

test loss of a classifier w is given by

$$\begin{aligned}&\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathcal{N}}}[\mathbb{1}[y(x-w)<0]]\\=&\frac{1}{2}+\frac{1}{2}\left(\Phi\left(\frac{w-\mu}{\sigma}\right)-\Phi\left(\frac{w+\mu}{\sigma}\right)\right),\end{aligned}\tag{9}$$

where Φ is the CDF of the standard normal distribution. In Section 4.2, we explain that the optimal classifier in hindsight is the one that is closest to 0 among the interval of classifiers.

Fig. 4a and Fig. 4b illustrate the test loss versus the size of the training dataset under the agnostic tiebreak and the optimal tiebreak in hindsight. We set $\mu = \sigma = 1$ and use the same set of values for ε for both tiebreaking methods. We have three observations. First, the generalization error is increasing in n when ε is larger than the signal strength. This confirms the existence of the strong adversary regime under the 0-1 loss. Second, for small enough ε (e.g. $\varepsilon \leq 0.5$), the generalization error is decreasing in n (more precisely after $n = 3$), thus also confirming a weak adversary regime. For the medium adversary where ε is in between 0.7 and 1.0, the curve has an increasing stage followed by a decreasing stage, which is very similar to what we see in Fig. 2b.

5.2 SUPPORT VECTOR MACHINE

We study the soft-margin support vector machine with hinge loss. The dimension d equals 2 and the data is generated as $y \sim \mathrm{Unif}(\{\pm 1\})$ and $X \sim \mathcal{N}(y\mu, I)$ where $\mu = (1, 1)^\top$. We consider the common setting of hinge loss with ℓ_2 penalty, under which the robust classifier is defined by

$$w_n^{\text{rob}} \in \arg\min_w \left(\frac{1}{n}\sum_{i=1}^n \ell_i(w)\right) + \frac{1}{2}\lambda\|w\|_2^2, \tag{10}$$

where

$$\ell_i(w) := \max_{\|x_i'-x_i\|_\infty \leq \varepsilon} \max\left\{0,\ 1 - y_i\left(\langle w, x_i'\rangle - b\right)\right\}.$$

The results are shown in Fig. 4c and Fig. 4d. The standard test loss (the y-axis in Fig. 4c and Fig. 4d) of the robust classifier is given by

$$\mathbb{E}_{(x,y)} \max\left\{0,\ 1 - y\left(\langle w, x\rangle - b\right)\right\},$$

where the penalty term is not included. The robust classifier w_n^{rob} is solved for by optimizing (10) which is convex in w using gradient descent.

We find that for small ε the standard test loss keeps decreasing, while for large ε it keeps increasing. The curves reveal a transition from the weak to the strong regime as ε grows, and such transition occurs when ε is in between 0.5 and 0.7. Note that at $\varepsilon = 0.7$, the test loss increases even though the strength of the adversary is still weaker than the signal level. This may indicate that for more complicated models (such as SVMs), even relatively weaker adversaries can result in situations where more data always increases the test loss.

5.3 LINEAR REGRESSION

Besides classification problems, we also identify similar phenomenon in one-dimensional linear regression $y = w^*x + e$ where $e \sim \mathcal{N}(0, 1)$ with squared loss $\ell(x, y; w) = (y - wx)^2$. Fig. 5 shows experimental results for linear regression. Given the coefficient w trained on n data points, the test loss is

$$\begin{aligned}L_n &= \mathbb{E}_{x,y}[(y - wx)^2] = \mathbb{E}_{x,e}[(w^*x + e - wx)^2]\\&= \mathbb{E}((w^* - w)x)^2 + \mathbb{E}e^2 = (w - w^*)^2\mathbb{E}x^2 + \mathbb{E}e^2.\end{aligned}$$

Therefore, in Fig. 5, we report the scaled test loss given by

$$\tilde{L}_n = \frac{(L_n - \mathbb{E}e^2)}{\mathbb{E}x^2} = (w - w^*)^2.$$

We use two different distributions for x: the Gaussian distribution $\mathcal{N}(0, 1)$ and the shifted Poisson distribution $\mathrm{Poisson}(5) + 1$. We add 1 to the outcome of $\mathrm{Poisson}(5)$ in order to guarantee a nonzero x. In both Gaussian and Poisson cases, we observe weak and strong regimes. When the perturbation strength ε is less than a threshold, it falls into the weak regime where the (scaled) test loss is reduced with more training data. When ε exceeds the threshold, it exhibits the strong regime where more data hurts the (scaled) test loss. However, the threshold is remarkably different for these two distributions. The threshold for $\mathcal{N}(0, 1)$ is between 0.6 and 0.8, while it resides between 4.0 and 7.0 for $\mathrm{Poisson}(5) + 1$. This observation suggests that the Poisson data distribution appears to be more robust to adversarial perturbation. A possible explanation could be that the distribution $\mathrm{Poisson}(5) + 1$ is supported on positive integers so the minimum distance between data points is 1, while there is no such minimum distance for data points following $\mathcal{N}(0, 1)$ and it becomes increasingly crowded as we have more data. As a result, adversarial perturbation has a stronger influence on Gaussian data.

6 CONCLUSION

The goal of adversarial training is to produce robust models that provide protection against attacks that make perturbations to the data at test time. While protection against such attacks is undoubtedly important, we still want our robust models to perform well on unperturbed data. However, our theoretical work shows the existence of scenarios in which current robust models do not achieve the desired low generalization error on both datasets simultaneously. This is a theoretical evidence of the gap between the standard accuracy of standard and robust models. Our findings suggest that the current adversarial training framework may not be ideal and that new ideas may be required to develop models that can reliably perform well on both perturbed and unperturbed test sets. It would also be interesting to develop theoretical results for the sample-wise non-monotonicity of more complicated models such as the neural nets.

Acknowledgements

Amin Karbasi is supported by NSF (IIS-1845032) and ONR (N00014-19-1-2406).

References

Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. Adversarial learning guarantees for linear hypotheses and neural networks. In *International Conference on Machine Learning*, pages 431–441. PMLR, 2020.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

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

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019b.

Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems*, pages 7496–7508, 2019.

Sebastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, pages 831–840, 2019.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE, 2018.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 11192–11203, 2019.

Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. Multiple descent: Design your own generalization curve. *arXiv preprint arXiv:2008.01036*, 2020a.

Lin Chen, Yifei Min, Mingrui Zhang, and Amin Karbasi. More data can expand the generalization gap between adversarially robust and standard models. In *International Conference on Machine Learning*, pages 1670–1680. PMLR, 2020b.

Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. Pac-learning in the presence of adversaries. In *Advances in Neural Information Processing Systems*, pages 230–241, 2018.

Dimitrios I Diochnos, Saeed Mahloujifar, and Mohammad Mahmoody. Lower bounds for adversarially robust pac learning. *arXiv preprint arXiv:1906.05815*, 2019.

Elvis Dohmatob. Generalized no free lunch theorem for adversarial robustness. In *International Conference on Machine Learning*, pages 1646–1654, 2019.

Ruiqi Gao, Tianle Cai, Haochuan Li, Cho-Jui Hsieh, Liwei Wang, and Jason D Lee. Convergence of adversarial training in overparametrized neural networks. In *Advances in Neural Information Processing Systems*, pages 13009–13020, 2019.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *stat*, 1050:20, 2015.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

Ruitong Huang, Bing Xu, Dale Schuurmans, and Csaba Szepesvári. Learning with a strong adversary. *arXiv preprint arXiv:1511.03034*, 2015.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *NeurIPS 2019*, pages 125–136, 2019.

Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, pages 2034–2078. PMLR, 2020.

Justin Khim and Po-Ling Loh. Adversarial risk bounds for binary classification via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 36–42. IEEE, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Paras Lakhani and Baskaran Sundaram. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2):574–582, 2017.

Shangbang Long, Yushuo Guan, Kaigui Bian, and Cong Yao. A new perspective for flexible feature gathering in scene text recognition via character anchor pooling. In *ICASSP 2020*, pages 2458–2462. IEEE, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR 2018*. OpenReview.net, 2018.

Hongyuan Mei, Sheng Zhang, Kevin Duh, and Benjamin Van Durme. Halo: Learning semantics-aware representations for cross-lingual information extraction. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 142–147, 2018.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Omar Montasser, Surbhi Goel, Ilias Diakonikolas, and Nathan Srebro. Efficiently learning adversarially robust halfspaces with noise. In *ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, pages 7010–7021. PMLR, 2020.

Amir Najafi, Shin-ichi Maeda, Masanori Koyama, and Takeru Miyato. Robustness to adversarial perturbations in learning from incomplete data. In *Advances in Neural Information Processing Systems*, pages 5541–5551, 2019.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2019.

Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*, 2016.

Apapan Pumsirirat and Liu Yan. Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine. *International Journal of advanced computer science and applications*, 9(1):18–25, 2018.

Muni Sreenivas Pydi and Varun Jog. Adversarial risk via optimal transport and optimal couplings. In *International Conference on Machine Learning*, pages 7814–7823. PMLR, 2020.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. In *ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena*, 2019.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *International Conference on Machine Learning*, 2020.

Sebastian Ramos, Stefan Gehrig, Peter Pinggera, Uwe Franke, and Carsten Rother. Detecting unexpected obstacles for self-driving cars: Fusing deep learning and geometric modeling. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1025–1032. IEEE, 2017.

Shibani Santurkar, Andrew Ilyas, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Image synthesis with a single (robust) classifier. In *Advances in Neural Information Processing Systems*, pages 1260–1271, 2019.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pages 5014–5026, 2018.

Uri Shaham, Yutaro Yamada, and Sahand Negahban. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 307:195–204, 2018.

Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Improving the generalization of adversarial training with domain adaptation. In *International Conference on Learning Representations*, 2019.

David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6976–6987, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *ICLR 2014*, 2014.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *ICLR*, 2019.

Tao Wang, David J Wu, Adam Coates, and Andrew Y Ng. End-to-end text recognition with convolutional neural networks. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 3304–3308. IEEE, 2012.

Colin Wei and Tengyu Ma. Improved sample complexities for deep networks and robust classification via an all-layer margin. *arXiv preprint arXiv:1910.04284*, 2019.

Jun Xiao, Yuanxing Zhang, Kaigui Bian, Guopeng Zhou, and Wei Yan. Denxfpn: Pulmonary pathologies detection based on dense feature pyramid networks. In *ICASSP*, pages 1234–1238. IEEE, 2019.

Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *ICML*, pages 7085–7094, 2019.

Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. In *Advances in Neural Information Processing Systems*, pages 227–238, 2019a.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482, 2019b.

Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *ICML*, 2020.

Tianyuan Zhang and Zhanxing Zhu. Interpreting adversarially trained convolutional neural networks. In *International Conference on Machine Learning*, pages 7502–7511, 2019.