More Data Can Expand the Generalization Gap Between Adversarially Robust and Standard Models

Lin Chen [*1] **Yifei Min** [*2] **Mingrui Zhang** [2] **Amin Karbasi** [1]

Abstract

Despite remarkable success in practice, modern machine learning models have been found to be susceptible to adversarial attacks that make human-imperceptible perturbations to the data, but result in serious and potentially dangerous prediction errors. To address this issue, practitioners often use adversarial training to learn models that are robust against such attacks at the cost of higher generalization error on unperturbed test sets. The conventional wisdom is that more training data should shrink the gap between the generalization error of adversarially-trained models and standard models. However, we study the training of robust classifiers for both Gaussian and Bernoulli models under ℓ_∞ attacks, and we prove that more data may actually increase this gap. Furthermore, our theoretical results identify if and when additional data will finally begin to shrink the gap. Lastly, we experimentally demonstrate that our results also hold for linear regression models, which may indicate that this phenomenon occurs more broadly.

1. Introduction

As modern machine learning models continue to gain traction in the real world, a wide variety of novel problems have come to the forefront of the research community. One particularly important challenge has been that of adversarial attacks (Szegedy et al., 2014; Goodfellow et al., 2015; Kos et al., 2018; Carlini & Wagner, 2018). To be specific, given a model with excellent performance on a standard data set, one can add small perturbations to the test data that can fool the model and cause it to make wrong predictions. What is more worrying is that these small perturbations can possibly be designed to be imperceptible to human beings, which raises concerns about potential safety issues and risks, especially when it comes to applications such as autonomous vehicles where human lives are at stake.

The problem of adversarial robustness in machine learning models has been explored from several different perspectives since its discovery. One direction has been to propose attacks that challenge these models and their training procedures (Gu & Rigazio, 2015; Moosavi-Dezfooli et al., 2016; Papernot et al., 2016; Carlini & Wagner, 2017; Athalye et al., 2018). In response, there have been works that propose more robust training techniques that can defend against these adversarial attacks (He et al., 2017; Raghunathan et al., 2018a;b; Shaham et al., 2018; Weng et al., 2018; Wong & Kolter, 2018; Zhang et al., 2018; Cohen et al., 2019; Lecuyer et al., 2019; Stutz et al., 2020). For robust training, one promising approach is to treat the problem as a minimax optimization problem, where we try to select model parameters that minimize the loss function under the strongest feasible perturbations (Xu & Mannor, 2012; Madry et al., 2018). Overall, adversarial training may be computationally expensive (Bubeck et al., 2019; Nakkiran, 2019), but it can lead to enhanced resistance towards adversarially modified inputs.

Although adversarially robust models tend to outperform standard models when it comes to perturbed test sets, recent studies have found that such robust models are also likely to perform worse on standard (unperturbed) test sets (Raghunathan et al., 2019; Tsipras et al., 2019). We refer to the difference in test loss on unperturbed test sets as the cross generalization gap. This paper focuses on the question of whether or not this gap can be closed.

Theoretical work by Schmidt et al. (2018) has shown that adversarial models require far more data than their standard counterparts to reach a certain level of test accuracy. This supports the general understanding that adversarial training is harder than standard training, as well as the conventional wisdom that more data helps with generalization. However, when it comes to the cross generalization gap, things may not be so simple.

[*]Equal contribution [1]Department of Electrical Engineering, Yale University [2]Department of Statistics and Data Science, Yale University. Correspondence to: Lin Chen <lin.chen@yale.edu>, Yifei Min <yifei.min@yale.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

In this paper, we identify two regimes during the adversarial training process. In one regime, more training data eventually helps to close the cross generalization gap, as expected. In the other regime, the gap will surprisingly continue to grow as more data is used in training. The data distribution and the strength of the adversary determine the regime and the existence of the two regimes indicates a fundamental phase transition in adversarial training.

1.1. Our Contributions

In our analysis of the cross generalization gap, we assume the robust model is trained under ℓ_∞ constrained perturbations. We study two classification models including a Gaussian model and a Bernoulli model, as well as a simple linear regression model.

For the Gaussian model, we theoretically prove that during the training of a robust classifier there are two possible regimes that summarize the relation between the cross generalization gap and the training sample size (see Theorem 1). More specifically, let n denote the number of training data points. Suppose the perturbation that the adversary can add is constrained to the ℓ_∞ ball of radius ε. In the strong adversary regime (i.e. large ε compared to the signal strength of the data), the gap always increases and has an infinite data limit.

In contrast, in the weak adversary regime, there exists a critical point that marks the boundary between two stages. For all n less than this threshold, we have the increasing stage where the gap monotonically increases. Beyond this threshold, we will eventually reach another stage where the gap strictly decreases. It is important to note that, even in the weak adversary regime, it is possible to make this threshold arbitrarily large, which means adding data points will always expand the cross generalization gap.

For the Bernoulli model, we show similar results (see Theorem 3). Although the curve for the cross generalization gap will be oscillating (see Fig. 1b), we prove that it manifests in a general increasing or decreasing trend. We further explore a simple one-dimensional linear regression and experimentally verify that the phase transition also exists.

The primary implication of our work is that simply adding more data will not always be enough to close the cross generalization gap. Therefore, fundamentally new ideas may be required if we want to be able to train adversarially robust models that do not sacrifice accuracy on unperturbed test sets.

2. Related Work

There is an existing body of work studying adversarially robust models and their generalization. We briefly discuss some of the papers that are most relevant to our work.

Trade-off between robustness and standard accuracy What initially motivated our work is the experimental finding that standard accuracy and adversarial robustness can sometimes be incompatible with each other (Papernot et al., 2018; Tsipras et al., 2019). These works empirically show that using more data for adversarial training might decrease the standard accuracy. Additionally, this decline becomes more obvious when the radius of perturbation ε increases. This causes the cross generalization gap between robust and standard models. The side effect of a large perturbation has also been studied by Dohmatob (2019) who shows that it is possible to adversarially fool a classifier with high standard accuracy if ε is large. Ilyas et al. (2019) explore the relation between the perturbation ε and the features learned by the robust model. Their results suggest that a larger ε tends to add more weight onto non-robust features and consequently the model may miss useful features which should be learned under standard setting. Diochnos et al. (2018) consider both error region setting and study the classification problem where data is uniformly distributed over $\{0,1\}^d$. They show that under this ℓ_0 perturbation setting the adversary can fool the classifier into having arbitrarily low accuracy with at most $\varepsilon = O(\sqrt{d})$ perturbation. Zhang et al. (2019) theoretically study the trade-off between robustness and standard accuracy from a perspective of decomposition. More specifically, they decompose the robust error into a standard error and a boundary error that would be affected by the perturbation. Their decomposition further leads to a new design of defense. Empirically, to deal with the reduction in the standard accuracy, Stutz et al. (2019) show that if the perturbation is not large enough to push data points across the decision boundary and the resulting adversarial examples still stay within their true decision region, then the adversarial training with such examples can boost generalization. Zhang et al. (2020) also propose training on specifically chosen adversarial examples to reduce the drop in the standard accuracy. Brittleness/robustness of Bayesian Inference is studied by Owhadi & Scovel (2016); Owhadi et al. (2015a;b); Owhadi & Scovel (2017).

In a concurrent and independent work, Raghunathan et al. (2020) performed a finite-sample analysis of the trade-off for a linear regression model. They also leveraged the recently proposed robust self-training estimator (Carmon et al., 2019; Najafi et al., 2019) in order to mitigate the robust error without sacrificing the standard error. They focused on a regression problem on the original training dataset augmented with perturbed examples and investigated a regime where the optimal predictor has zero standard and robust error. This paper studied a classification problem and our analysis covers both weak and strong regimes.

Sample complexity for generalization The generalization of adversarially robust models has different properties from the standard ones, especially in sample complexity. Schmidt et al. (2018) study Gaussian mixture models in d-dimensional space and show that for the standard model only a constant number of training data points is needed, while for the robust model under ℓ_∞ perturbation a training dataset of size $\Omega(d)$ is required. Their work is in a different direction to ours: their main result focuses on dimension-dependent bounds for sample complexity, while we quantify the effect of the amount of training data on adversarial generalization and we prove the existence of a phase transition under two binary classification models. Bubeck et al. (2019) analyze the computational hardness in training a robust classifier in the statistical query model. They prove that for a binary classification problem in d dimensions, one needs polynomially (in d) many queries to train a standard classifier while exponentially many queries to train a robust one. Garg et al. (2020) consider a setting where the adversary has limited computational power and show that there exist learning tasks that can only be robustly solved when faced with such limited adversaries. Yin et al. (2019) and Khim & Loh (2018) prove generalization bounds for linear classifiers and neural networks via Rademacher complexity. In addition, Yin et al. (2019) show the adversarial Rademacher complexity is always no less than the standard one and is dimension-dependent. Montasser et al. (2019) show the widely used uniform convergence of empirical risk minimization framework, or more generally, any proper learning rule, might not be enough for robust generalization. They prove the existence of a hypothesis class where any proper learning rule gives poor robust generalization accuracy under the PAC-learning setting, while improper learning can robustly learn any class. Cullina et al. (2018) study generalization under the PAC-learning setting and prove a polynomial upper bound for sample complexity that depends on a certain adversarial VC-dimension. Diochnos et al. (2020) study PAC-learning under the error region setting and prove a lower bound for sample complexity that is exponential in the input dimension.

Other relevant work Bhagoji et al. (2019) use optimal transport to derive lower bounds for the adversarial classification error. For a binary classification problem, they prove a relation between the best possible adversarial robustness and the optimal transport between the two distributions under a certain cost. Another line of work analyzes adversarial examples via concentration of measure and show that their existence is inevitable under certain conditions (Gilmer et al., 2018; Fawzi et al., 2018; Shafahi et al., 2018; Mahloujifar et al., 2019).

3. Preliminaries

3.1. Notation

We use the shorthand $[d]$ to denote the set $\{1, 2, \ldots, d\}$ for any positive integer d. We use $\mathcal{N}(\mu, \Sigma)$ to denote the multivariate Gaussian distribution with mean vector μ and covariance matrix Σ.

If $u, v \in \mathbb{R}^d$ are two d-dimensional vectors, the j-th component of u is denoted by $u(j)$. The inner product of u and v is denoted by $\langle u, v\rangle$. If A is a positive semi-definite matrix, let the semi-norm induced by A be $\|u\|_A = \sqrt{u^\top A u}$. Let $B_u^\infty(\varepsilon)$ denote the ℓ_∞ ball centered at u and with radius ε, i.e., $B_u^\infty(\varepsilon) = \{v \in \mathbb{R}^d : \|u - v\|_\infty \le \varepsilon\}$. In our problem setup in Section 3.2, the ball B_u^∞ is the set of allowed perturbed vectors for the adversary, where ε is the perturbation budget. We define the Heaviside step function H to be

$$H(x) = \begin{cases} 1, & \text{for } x > 0\,; \\ 1/2, & \text{for } x = 0\,; \\ 0, & \text{for } x < 0\,. \end{cases}$$

3.2. Problem Setup

Suppose that the data (x, y) is drawn from an unknown distribution $\mathcal{D}$, where x is the input and y is the label. For example, in a classification problem, we have $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$; in a regression problem, we have $(x, y) \in \mathbb{R}^d \times \mathbb{R}$. Given a model parameter $w \in \Theta \subseteq \mathbb{R}^p$ and a data point (x, y), the loss of the model parameterized by w on the data point (x, y) is denoted by $\ell(x, y; w)$.

The training dataset $D_{\text{train}} = \{(x_i, y_i)\}_{i=1}^n$ consists of n data points sampled i.i.d. from the distribution $\mathcal{D}$. Given the training dataset with size n, we respectively define the optimal standard and robust models trained on D_{train} by

$$\begin{aligned} w_n^{\text{std}} &= \arg\min_{w\in\Theta} \frac{1}{n}\sum_{i=1}^n \ell(x_i, y_i; w)\,, \\ w_n^{\text{rob}} &= \arg\min_{w\in\Theta} \frac{1}{n}\sum_{i=1}^n \max_{\tilde{x}_i \in B_{x_i}^\infty(\varepsilon)} \ell(\tilde{x}_i, y_i; w)\,. \end{aligned} \tag{1}$$

The optimal standard model w^{std} is the minimizer of the total training loss $\frac{1}{n}\sum_{i=1}^n \ell(x_i, y_i; w)$. In the definition of the optimal robust model w^{rob}, we take into consideration the adversarial training for each data point, i.e., the inner maximization $\max_{\tilde{x}_i \in B_{x_i}^\infty(\varepsilon)} \ell(\tilde{x}_i, y_i; w)$. We assume that the adversary is able to perturb each data item x_i within an ℓ_∞ ball centered at x_i and with radius ε. The best robust model is the minimizer of the total training loss with adversarial training. Note that both w^{std} and w^{rob} are functions of the training dataset and thereby also random variables.

If we have a model parametrized w and the test dataset $D_{\text{test}} = \{(x_i', y_i')\}_{i=1}^{n'}$ consists of n' data points sampled

i.i.d. from $\mathcal{D}$, the test loss of w is given by

$$L_{\text{test}}(w) = \mathbb{E}\left[\frac{1}{n'}\sum_{i=1}^{n'} \ell(x_i', y_i'; w)\right] = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(x, y; w)\right] .$$

Additionally, we define the cross generalization gap g_n between the standard and robust classifiers by

$$\begin{aligned} g_n &= \mathbb{E}_{\{(x_i,y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim}\mathcal{D}}\left[L_{\text{test}}(w^{\text{rob}}) - L_{\text{test}}(w^{\text{std}})\right] \\ &= \mathbb{E}_{\{(x_i,y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim}\mathcal{D}}\left[\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(x, y; w^{\text{rob}})]\right. \\ &\quad \left. -\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(x, y; w^{\text{std}})]\right] . \end{aligned}$$

4. Classification

In this section, we study a binary classification problem, where we have each data point $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$. For any model parameter $w \in \mathbb{R}^d$, we consider the loss function $\ell(x, y; w) = -y\langle w, x\rangle$ (Yin et al., 2019; Khim & Loh, 2018). The parameter w is constrained on the ℓ^∞ ball $\Theta = \{w \in \mathbb{R}^d \mid \|w\|_\infty \le W\}$, where W is some positive real number. Under this setup, the best standard and robust classifier are given as follows.

$$\begin{aligned} w_n^{\text{std}} &= \operatorname*{arg\,min}_{\|w\|_\infty\le W} \frac{1}{n}\sum_{i=1}^n -y_i\langle w, x_i\rangle \\ &= \operatorname*{arg\,max}_{\|w\|_\infty\le W} \sum_{i=1}^n y_i\langle w, x_i\rangle , \\ w_n^{\text{rob}} &= \operatorname*{arg\,min}_{\|w\|_\infty\le W} \frac{1}{n}\sum_{i=1}^n \max_{\tilde{x}_i\in B_{x_i}^\infty(\varepsilon)} (-y_i\langle w, \tilde{x_i}\rangle) \\ &= \operatorname*{arg\,max}_{\|w\|_\infty\le W} \sum_{i=1}^n \min_{\tilde{x}_i\in B_{x_i}^\infty(\varepsilon)} y_i\langle w, \tilde{x_i}\rangle . \end{aligned} \tag{2}$$

The cross generalization gap g_n between the standard and robust classifiers is given by

$$g_n = \mathbb{E}_{\{(x_i,y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim}\mathcal{D}}\left[\mathbb{E}_{(x,y)\sim\mathcal{D}}[y\langle w^{\text{std}}, x\rangle] - \mathbb{E}_{(x,y)\sim\mathcal{D}}[y\langle w^{\text{rob}}, x\rangle]\right] . \tag{3}$$

In this paper, we investigate how the cross generalization gap g_n evolves with the amount of data. Intuitively, one might conjecture that the gap should satisfy the following properties:

(a) First, the gap should always be non-negative. This means that the robust classifier incurs a larger test (generalization) loss than the standard classifier, as there is no free lunch and robustness in adversarial training would compromise generalization performance.

(b) Second, more training data would close the gap gradually; in other words, the gap would be decreasing with respect to the size of the training dataset.

(c) Third, in the infinite data limit (i.e., when the size of the training dataset tends to infinity), the cross generalization gap would eventually tend to zero.

Our study corroborates (a) but denies (b) and (c) in general. The implication of this is not only that current adversarial training techniques sacrifice standard accuracy in exchange for robustness, but that simply adding more data may not solve the problem.

4.1. Gaussian Model

The Gaussian model is specified as follows. Let $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$ obey the distribution such that $y \sim \text{Unif}(\{\pm 1\})$ and $x \mid y \sim \mathcal{N}(y\mu, \Sigma)$, where $\mu(j) \ge 0$ for $\forall j \in [d]$ and $\Sigma = \text{diag}(\sigma(1)^2, \sigma(2)^2, \ldots, \sigma(d)^2)$. We denote this distribution by $(x, y) \sim \mathcal{D}_{\text{Gau}}$.

Theorem 1 (Gaussian model, **proof in Appendix A**). *Given i.i.d. training data $(x_i, y_i) \sim \mathcal{D}_{\text{Gau}}$ with n data points, if we define the standard and robust classifier as in* (2) *(denoted by w^{std} and w^{rob}, respectively) and define the cross generalization gap g_n as in* (3)*, we have*

(a) $g_n \ge 0 \ \forall\, n \ge 1$;

(b) The infinite data limit equals

$$\lim_{n\to\infty} g_n = 2W \sum_{j\in[d]:\mu(j)>0} \mu(j) H\left(\frac{\varepsilon}{\mu(j)} - 1\right) ,$$

where H is the Heaviside step function defined in Section 3.1;

(c) If $\varepsilon < \min_{j\in[d]:\mu(j)>0} \mu(j)$, g_n is strictly increasing in n when

$$n < \min_{\substack{j\in[d]:\\ \mu(j)>0}} \max\left\{\frac{3}{2}, 2\log\frac{1}{1-\varepsilon/\mu(j)}\right\}\left(\frac{\sigma(j)}{\mu(j)}\right)^2 ,$$

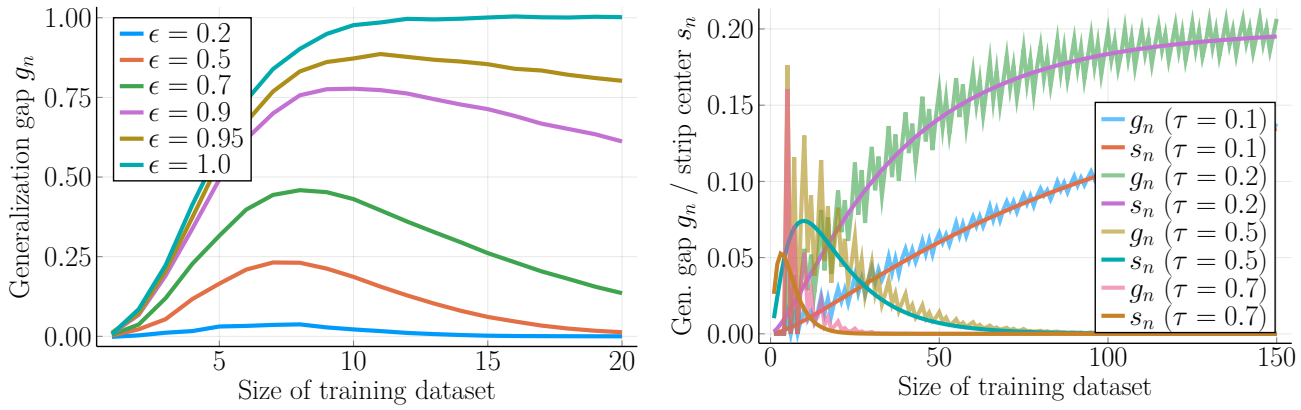
and it is strictly decreasing in n when

$$n \ge \max_{\substack{j\in[d]:\\ \mu(j)>0}} \left(K_0 + 2\log\frac{1}{1-\varepsilon/\mu(j)}\right)\left(\frac{\sigma(j)}{\mu(j)}\right)^2 ,$$

where K_0 is a universal constant.

(d) If $\varepsilon > \|\mu\|_\infty$, g_n is strictly increasing for all $n \ge 1$.

Part (a) of Theorem 1 states that the generalization of the robust classifier is never better than the standard one. Part (b) quantifies the size of the gap as the size of the training dataset n goes to infinity. The main implication here is that the gap will always converge to some finite limit, which may be zero if the strength of the adversary ϵ is small enough.



(a) Gaussian model

(b) Bernoulli model

Figure 1: Cross generalization gap g_n (and strip center s_n for the Bernoulli model) vs. the size of the training dataset.

Parts (c) and (d) describe the two different possible regimes. Part (c) states that if the strength of the adversary is not too large, then there will be two stages: an initial stage where the cross generalization gap is strictly increasing in n, followed by a secondary stage where the gap is strictly decreasing in n. On the other hand, part (d) states that a large ϵ will result in a cross generalization gap that is strictly increasing (but still tending towards some finite limit).

In order to better describe and visualize the implications of Theorem 1, we consider a special case where $\mu = (\mu_0, \ldots, \mu_0)$ and $\Sigma = \sigma_0^2 I$.

Corollary 2. *Assume that $W = 1$, $\mu(j) = \mu_0 \geq 0$, and $\sigma(j) = \sigma_0 > 0$ for all $j \in [d]$. The infinite data limit equals*

$$\lim_{n\to\infty} g_n = 2d\mu_0 H\left(\frac{\varepsilon}{\mu_0} - 1\right) = \begin{cases} 2d\mu_0, & \text{for } \frac{\varepsilon}{\mu_0} > 1\,; \\ d\mu_0, & \text{for } \frac{\varepsilon}{\mu_0} = 1\,; \\ 0, & \text{for } \frac{\varepsilon}{\mu_0} < 1\,. \end{cases}$$

If $\varepsilon < \mu_0$, we have g_n is strictly increasing when

$$n < \max\left\{\frac{3}{2}, 2\log\frac{1}{1-\varepsilon/\mu_0}\right\}\left(\frac{\sigma_0}{\mu_0}\right)^2,$$

and it is strictly decreasing when

$$n \geq \left(K_0 + 2\log\frac{1}{1-\varepsilon/\mu_0}\right)\left(\frac{\sigma_0}{\mu_0}\right)^2,$$

where K_0 is a universal constant. If $\varepsilon > \mu_0$, we have g_n is strictly increasing for all $n \geq 1$.

Corollary 2 is essentially a simplified version of parts (c) and (d) of Theorem 1 where we cleanly divide between a weak adversary regime and a strong adversary regime at a threshold $\varepsilon = \mu_0$.

We illustrate the cross generalization gap g_n vs. the size of the training dataset in Fig. 1a, where we set $W = d = \mu = 1$ and $\sigma = 2$. The curve $\varepsilon = 1$ belongs to the strong adversary regime, while the remaining curves belong to the weak adversary regime.

In the weak adversary regime, the evolution of g_n can be divided into two stages, namely the increasing and decreasing stages (part (c) of Theorem 1). The duration of the increasing stage is

$$\Theta\left(\left(\frac{\sigma_0}{\mu_0}\right)^2 \log\frac{1}{1-\varepsilon/\mu_0}\right).$$

This duration is controlled by the ratio ε/μ_0, as well as the reciprocal of the signal-to-noise ratio (SNR), i.e., $\frac{\sigma_0}{\mu_0}$. A larger SNR and an ε closer to μ_0 lead to a shorter increasing stage. It can be observed in Fig. 1a that for the curves with $\varepsilon = 0.2, 0.5, 0.7, 0.9, 0.95$, a larger ε results in a longer duration of the increasing stage.

After the increasing stage, the cross generalization gap will eventually begin to decrease towards some finite limit (given by part (b) of Theorem 1) if sufficient training data is provided. In addition, we would like to remark that the duration relies on the data and the strength of the adversary and could be potentially arbitrarily large; in other words, without full information about the true data distribution and the power of the adversary, one cannot predict when the increasing stage will terminate.

In the strong adversary regime, the cross generalization gap expands from the very beginning. In the infinite data limit, the gap approaches $d\mu_0$ if $\varepsilon = \mu_0$, and it approaches $2d\mu_0$ if $\varepsilon > \mu_0$.

4.2. Bernoulli Model

In this subsection, we investigate the Bernoulli model defined as follows. Let $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$ obey the distribution such that $y \sim \mathrm{Unif}(\{\pm 1\})$ and for $\forall j \in [d]$ indepen-

dently,

$$x(j) = \begin{cases} y \cdot \theta(j) & \text{with probability } \frac{1+\tau}{2}\,, \\ -y \cdot \theta(j) & \text{with probability } \frac{1-\tau}{2}\,, \end{cases}$$

where $\theta \in \mathbb{R}^d_{\geq 0}$ and $\tau \in (0,1)$. We denote this distribution by $(x,y) \sim \widetilde{\mathcal{D}}_{\text{Ber}}$.

The parameter τ controls the signal strength level. When $\tau = 0$ (lowest signal strength), $x(j)$ takes the value of $+\theta(j)$ or $-\theta(j)$ uniformly at random, irrespective of the label y. When $\tau = 1$ (highest signal strength), we have $x(j) = y \cdot \theta(j)$ almost surely.

We illustrate the cross generalization gap g_n vs. the size of the training dataset (denoted by n) in Fig. 1b, where we set $W = d = \theta = 1$ and $\varepsilon = 0.2$. We observe that all curves g_n oscillate around the other curves labeled s_n. Although the figure shows that the curves g_n are not monotone, they all exhibit a monotone trend, which is characterized by s_n.

As a result, we will not show that g_n is monotonically increasing or decreasing (as shown in Fig. 1b, it is not monotone). Alternatively, we will show that g_n resides in a strip centered around s_n and s_n displays (piecewise) monotonicity. Additionally, the height of the strip shrinks at a rate of $O\left(\frac{1}{\sqrt{n}}\right)$; in other words, it can be shown that

$$|g_n - s_n| \leq O\left(\frac{1}{\sqrt{n}}\right), \quad \forall n \geq 1\,.$$

Theorem 3 (Bernoulli model, **proof in Appendix B**). *Given i.i.d. training data $(x_i, y_i) \sim \mathcal{D}_{\text{Ber}}$ with n data points, if we define the standard and robust classifier (denoted by w^{std} and w^{rob}, respectively) as in* (2) *and define the cross generalization gap g_n as in* (3)*, we have*

(a) $g_n \geq 0$ for $\forall n \geq 1$;
(b) The infinite data limit equals

$$\lim_{n\to\infty} g_n = 2W\tau \sum_{j\in[d]:\theta(j)>0} \theta(j) H\left(\frac{\varepsilon}{\theta(j)\tau} - 1\right),$$

where H is the Heaviside step function defined in Section 3.1.

Furthermore, there exists a positive constant $C_0 \leq \frac{\sqrt{10}+3}{6\sqrt{2\pi}} \approx 0.4097$ and a sequence s_n such that $|g_n - s_n| \leq \frac{8C_0 W\tau\|\theta\|_1(\tau^2+1)}{\sqrt{n}\sqrt{1-\tau^2}}$ and

(c) If $\frac{\varepsilon}{\tau} < \min_{j\in[d]:\theta(j)>0}\theta(j)$, s_n is strictly increasing in n when

$$n < \left(\frac{1}{\tau^2} - 1\right) \max\left\{\frac{3}{2}, 2 \min_{\substack{j\in[d]:\\ \theta(j)>0}} \log\frac{1}{1-\frac{\varepsilon}{\theta(j)\tau}}\right\}$$

and strictly decreasing in n when

$$n \geq \left(\frac{1}{\tau^2} - 1\right)\left(K_0 + 2 \max_{\substack{j\in[d]:\\ \theta(j)>0}} \log\frac{1}{1-\frac{\varepsilon}{\theta(j)\tau}}\right),$$

where K_0 is a universal constant;
(d) If $\frac{\varepsilon}{\tau} \geq \|\theta\|_\infty$, s_n is strictly increasing for all $n \geq 1$.

Again, to explain the implications of Theorem 3, we explore the following special case where $W = 1$ and $\theta = (\theta_0, \ldots, \theta_0)$.

Corollary 4. *Assume $W = 1$ and that $\theta(j) = \theta_0 > 0$ holds for all $j \in [d]$. The infinite data limit equals*

$$\begin{aligned} \lim_{n\to\infty} g_n &= 2\tau d\theta_0 H\left(\frac{\varepsilon}{\theta_0\tau} - 1\right) \\ &= \begin{cases} 2\tau d\theta_0, & \text{for } \varepsilon > \theta_0\tau\,; \\ \tau d\theta_0, & \text{for } \varepsilon = \theta_0\tau\,; \\ 0, & \text{for } \varepsilon < \theta_0\tau\,. \end{cases} \end{aligned} \tag{4}$$

If $\varepsilon < \theta_0\tau$, s_n is strictly increasing in n when

$$n < \left(\frac{1}{\tau^2} - 1\right) \max\left\{\frac{3}{2}, 2\log\frac{1}{1-\varepsilon/(\theta_0\tau)}\right\},$$

and it is strictly decreasing when

$$n \geq \left(\frac{1}{\tau^2} - 1\right)\left(K_0 + 2\log\frac{1}{1-\varepsilon/(\theta_0\tau)}\right),$$

where K_0 is a universal constant. If $\varepsilon \geq \theta_0\tau$, s_n is strictly increasing for all $n \geq 1$.

Similar to the Gaussian model, there also exist two regimes. One is the weak adversary regime where $\varepsilon < \theta_0\tau$, while the other is the strong adversary regime where $\varepsilon \geq \theta_0\tau$. Recall that in Fig. 1b, we set $W = d = \theta = 1$ and $\varepsilon = 0.2$. Therefore the values $\tau = 0.1$ and $\tau = 0.2$ lie in the strong adversary regime, while the values $\tau = 0.5$ and $\tau = 0.7$ belong to the weak adversary regime.

In the weak adversary regime, the critical point is when

$$n \approx \Theta\left(\left(\frac{1}{\tau^2} - 1\right)\log\frac{1}{1-\varepsilon/(\theta_0\tau)}\right). \tag{5}$$

Before this critical point, the strip center s_n that the cross generalization gap g_n oscillates around is strictly increasing; it is strictly decreasing after the critical point and eventually vanished as $n \to \infty$. Note that when $\tau \to 0$, both terms ($\left(\frac{1}{\tau^2} - 1\right)$ and $\log\frac{1}{1-\varepsilon/(\theta_0\tau)}$) in (5) blow up and thereby the increasing stage elongates infinitely. The increasing and decreasing stages of the weak adversary regime are confirmed by the two curves $\tau = 0.5$ and $\tau = 0.7$ in Fig. 1b.

In the strong adversary regime, the strip center s_n displays a similar trend as the cross generalization gap in the Gaussian model; i.e., it is strictly increasing from the very beginning (see the two curves $\tau = 0.1$ and $\tau = 0.2$ in Fig. 1b). Recall that under the Bernoulli model, the strong/weak adversary regime is determined by the ratio $\frac{\varepsilon}{\theta_0 \tau}$, while under the Gaussian model, it is determined by the ratio $\frac{\varepsilon}{\mu_0}$. Nevertheless, note that in the binary classification, $\theta_0 \tau$ is the mean (in one coordinate) of the positive class, just like μ_0 in the Gaussian scenario. These two ratios are thus closely related.

We would also like to remark that limits of g_n in Fig. 1b follow the theoretical results outlined in (4). In particular, if we are in the weak adversary regime, the limit of g_n always tends to 0. On the other hand, in the strong adversary regime, the limit is non-zero and proportional to τ.

4.3. Discussion

One common observation from Theorem 1 and Theorem 3 is that the duration of the increasing stage heavily depends on the ratio between ε and the coordinate-wise mean of the positive class (i.e. μ_0 and $\theta_0 \tau$). Note that the mean can be interpreted as half the distance between the centers of positive and negative classes in the space of x. Thus, another way to view this result is that if the strength of the adversary is relatively large compared to the distance between classes, then we will have a long increasing stage.

One interesting implication of this can be seen in regression vs. classification tasks. Intuitively, one might look at a regression task as a classification task with infinitely many classes. Therefore, depending on the distribution that x is sampled from, we could end up with a very small distance between class centers and thus we would expect a very long increasing stage.

5. Regression

In this section, we explore the problem of linear regression, where we have each data point $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ and the linear model is represented by a vector $w \in \mathbb{R}^d$. The loss function is defined by $\ell(x, y; w) = (y - \langle w, x \rangle)^2$.

We assume the following data generation process. First, we sample x_i from some distribution P_X. Given the fixed true model w^*, we set $y_i = \langle w^*, x_i \rangle + \delta$, where $\delta \sim \mathcal{N}(0, \sigma^2)$ is the Gaussian noise. The parameter space Θ is the entire $\mathbb{R}^d$.

Given the training dataset $D_{\text{train}} = \{(x_i, y_i)\}_{i=1}^n$, if we define $X = [x_1, \ldots, x_n]^\top$ and $y = [y_1, \ldots, y_n]^\top$, the best standard model has a closed form (Graybill, 1961):

$$w^{\text{std}} = (X^\top X)^{-1} X^\top y \,.$$

Observation 5 presents the form of the best robust model in the linear regression problem.

Observation 5 (Proof in Appendix C). *The best robust model in the linear regression problem is given by*

$$w_n^{\text{rob}} = \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \left(|y_i - \langle w, x_i \rangle| + \varepsilon \sum_{j=1}^d |w(j)| \right)^2 .$$

Observation 6 gives the form of the gap in the linear regression problem setting.

Observation 6 (Proof in Appendix D). *In the linear regression problem, the cross generalization gap equals*

$$g_n = \|w_n^{\text{rob}} - w^*\|^2_{\mathbb{E}_{x \sim P_X}[xx^\top]} - \|w_n^{\text{std}} - w^*\|^2_{\mathbb{E}_{x \sim P_X}[xx^\top]} \,.$$

Observation 6 shows that the cross generalization gap not only depends on the difference vectors $(w_n^{\text{rob}} - w^*)$ and $(w_n^{\text{std}} - w^*)$ but also the matrix $\mathbb{E}_{x \sim P_X}[xx^\top]$. This matrix weights each dimension of the difference vectors and thereby influences the cross generalization gap.

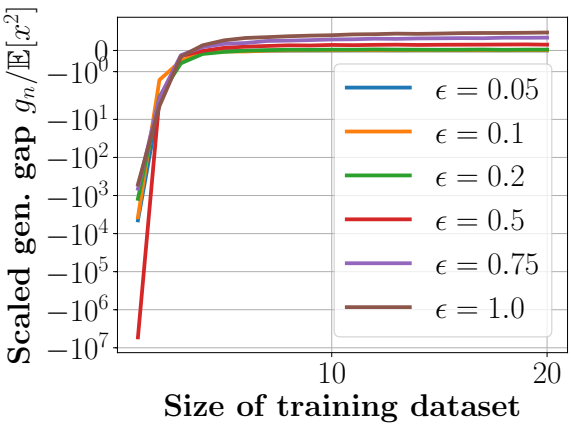
To avoid the complication incurred by the different weightings of the matrix $\mathbb{E}_{x \sim P_X}[xx^\top]$ across the dimensions, we investigate two one-dimensional linear regression problems ($d = 1$) with the data input x sampled from a standard normal distribution and a shifted Poisson distribution, respectively. To be specific, in the first study, we consider x sampled from the standard normal distribution $\mathcal{N}(0, 1)$. In the second study, the data input x is drawn from $\text{Poisson}(5) + 1$ (in order to avoid $x = 0$); in other words, $x - 1$ obeys the $\text{Poisson}(5)$ distribution. In both studies, we set the true model $w^* = 1$ and the noise obeys $\delta \sim \mathcal{N}(0, 1)$ (i.e., $\sigma^2 = 1$). In light of Observation 6, we obtain that if the linear regression problem is one-dimensional, the cross generalization gap equals

$$g_n = \mathbb{E}_{(x,y) \sim \mathcal{D}} \left((w_n^{\text{rob}} - w^*)^2 - (w_n^{\text{std}} - w^*)^2 \right) \mathbb{E}_{x \sim P_X}[x^2] \,.$$

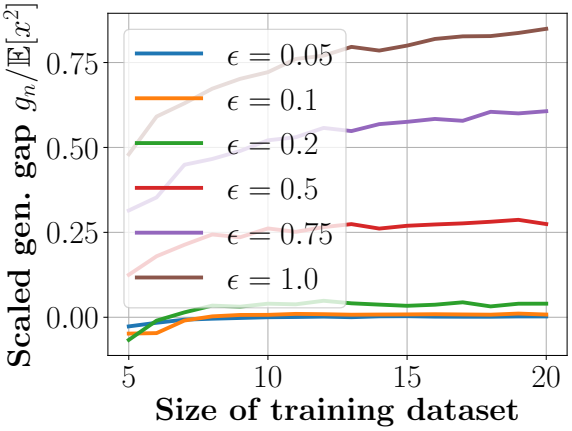
Since g_n is proportional to $\left((w_n^{\text{rob}} - w^*)^2 - (w_n^{\text{std}} - w^*)^2 \right)$ with $\mathbb{E}_{x \sim P_X}[x^2]$ being a constant, we call $g_n / \mathbb{E}_{x \sim P_X}[x^2]$ the *scaled cross generalization gap* and plot it against the size of the training dataset (denoted by n) in Fig. 2.

Fig. 2a shows the result for the first study with n ranging from 1 to 20. For a clear presentation, Fig. 2b provides a magnified plot for $5 \leq n \leq 20$.

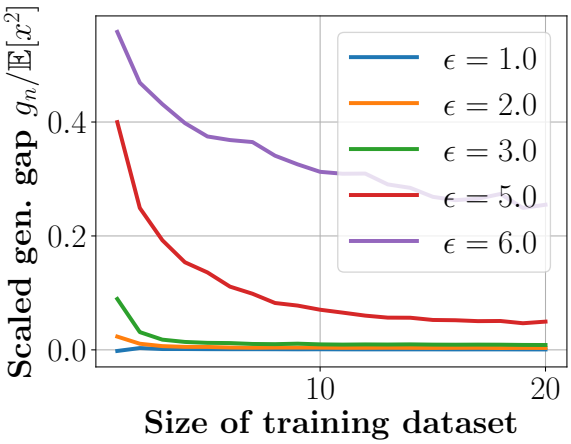
Our first observation is that in the Gaussian case, the cross generalization gap g_n always expands with more data, even if ε is as small as 0.05. This may be because if we sort n i.i.d. standard normal random variables $x_1, \ldots, x_n$ in ascending order and obtain $x_{\pi(1)} \leq x_{\pi(2)} \leq \cdots \leq x_{\pi(n)}$, the difference between two consecutive numbers (i.e., $x_{\pi(i+1)} - x_{\pi(i)}$) becomes smaller as n becomes larger. As we discussed in Section 4, the monotone trend of g_n is



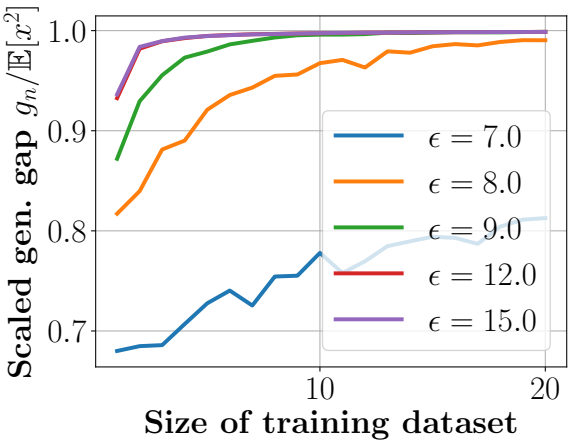
(a) $x \sim \mathcal{N}(0,1)$, $1 \le n \le 20$.



(b) $x \sim \mathcal{N}(0,1)$, $5 \le n \le 20$.



(c) $x \sim \text{Poisson}(5) + 1$, small ε.



(d) $x \sim \text{Poisson}(5) + 1$, large ε.

Figure 2: Scaled cross generalization gap $g_n/\mathbb{E}_{x\sim P_X}[x^2]$ vs. the size of the training dataset (denoted by n). First two plots correspond to x being sampled from the standard normal distribution $\mathcal{N}(0,1)$ and last two plots correspond to $\text{Poisson}(5)+1$. Each curve in a plot represents a different choice of ε.

determined by the ratio of ε to half the distance between the positive and negative classes. The ratio is $\frac{\varepsilon}{\mu_0}$ in the Gaussian model and it is $\frac{\varepsilon}{\theta_0 \tau}$ in the Bernoulli model. The regression problem may be viewed as a classification problem with infinitely many classes. The difference between two consecutive numbers is the analog of the distance between the means of difference classes. Since the difference reduces as n becomes larger (points are more densely situated), the ratio increases and therefore we observe a wider cross generalization gap.

Our second observation regarding the Gaussian data is that the cross generalization gap is (very) negative at the initial stage. In particular, when $n = 1$, the gap g_1 is between -10^6 and -10^7. The reason is that when $n = 1$, we have

$$\mathbb{E}\left[(w_1^{\text{std}} - w^*)^2\right] = \infty\,.$$

Because of the robustness, w^{rob} is more stablized and therefore $\mathbb{E}[\left(w_1^{\text{rob}} - w^*\right)^2]$ is finite. Since the cross generalization gap g_1 is proportional to their difference $\mathbb{E}[\left(w_1^{\text{rob}} - w^*\right)^2] - \mathbb{E}\left[(w_1^{\text{std}} - w^*)^2\right]$, the gap g_1 is indeed $-\infty$. We present a proof of $g_1 = -\infty$ in Theorem 7.

Theorem 7 (Proof in Appendix E). *In the one-dimensional linear regression problem, if $x_1 \sim \mathcal{N}(0,1)$, $\delta \sim \mathcal{N}(0,1)$, and $y_1 = w^* x + \delta$, the cross generalization gap g_1 with only one training data point is $-\infty$.*

Fig. 2c presents the result for the Poisson input with ε varying from 1.0 to 6.0. Fig. 2d illustrates the result corresponding to large ε that ranges from 7.0 to 15.0. We see two different regimes in Fig. 2c and Fig. 2d. Fig. 2c represents the weak adversary regime where the cross generalization gap shrinks with more training data. Fig. 2d represents the strong adversary regime in which the gap expands with more training data. Furthermore, given the same size of the

training dataset, the gap increases with ε.

The result for the Poisson input is in sharp contrast to the Gaussian input. It appears that for any small ε, the cross generalization gap will increase with more data in the Gaussian setting, as the real line becomes increasingly crowed with data points. In the Poisson setting, whilst the Poisson distribution is infinitely supported as well, the minimum distance between two different data points is one (recall that the Poisson distribution is supported on natural numbers). A weak adversary with a small ε is unable to drive the cross generalization gap into an increasing trend. Additionally, recalling that the mean of $\mathrm{Poisson}(5)+1$ is 6, the value $\varepsilon = 6$ exactly separates the weak and strong adversary regimes in these two figures. Note that all ε values in Fig. 2c are ≤ 6, while all those in Fig. 2d are > 6.

Unlike the Gaussian setting for linear regression, we never observe a negative cross generalization gap, even if $n = 1$. This observation supports our theoretical finding, which is summarized in Theorem 8.

Theorem 8 (**Proof in Appendix F**). *In the one-dimensional linear regression problem, if* $x_1 \sim \mathrm{Poisson}(\lambda) + 1$, $\delta \sim \mathcal{N}(0, 1)$, *and* $y_1 = w^* x + \delta$ *with* $|w^*| \geq 1$, *the cross generalization gap* g_1 *with only one training data point is non-negative, finite, and increases with* ε.

6. Conclusion

In this paper, we study the cross generalization gap between adversarially robust models and standard models. We analyze two classification models (the Gaussian model and the Bernoulli model), and we also explore the linear regression model. We theoretically find that a larger training dataset won't necessarily close the cross generalization gap and may even expand it. In addition, for the two classification models, we prove that the cross generalization gap is always non-negative, which indicates that current adversarial training must sacrifice standard accuracy in exchange for robustness.

For the Gaussian classification model, we identify two regimes: the strong adversary regime and the weak adversary regime. In the strong adversary regime, the cross generalization gap monotonically expands towards some non-negative finite limit as more training data is used. On the other hand, in the weak adversary regime, there are two stages: an increasing stage where the gap increases with the training sample size, followed by a decreasing stage where the gap decreases towards some finite non-negative limit. Broadly speaking, the ratio between the strength of the adversary and the distance between classes determines which regime we will fall under.

In the Bernoulli model, we also prove the existence of the weak and strong adversary regimes. The primary difference is that the cross generalization gap is oscillating instead of monotone. However, we also show that these oscillating curves have strip centers that display very similar behavior to the Gaussian curves.

Our findings are further validated by a study of the linear regression model, which experimentally exhibits similar behavior and may indicate that our results hold for an even broader class of models. The ultimate goal of adversarial training is to learn models that are robust against adversarial attacks, but do not sacrifice any accuracy on unperturbed test sets. The primary implication of our work is that this trade-off is provably unavoidable for existing adversarial training frameworks.

Acknowledgements

AK is partially supported by NSF (IIS-1845032), ONR (N00014-19-1-2406), and AFOSR (FA9550-18-1-0160). LC is supported by Google PhD Fellowship. We would like to thank Peter Bartlett, Hamed Hassani, Adel Javanmard, and Mohammad Mahmoody for their comments regarding the first version of the paper and thank Marko Mitrovic for his help in preparation of the paper.

References

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pp. 274–283, 2018.

Berry, A. C. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941.

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

Bubeck, S., Lee, Y. T., Price, E., and Razenshteyn, I. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, pp. 831–840, 2019.

Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 3–14. ACM, 2017.

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

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., and Liang, P. S. Unlabeled data improves adversarial ro-

bustness. In *Advances in Neural Information Processing Systems*, pp. 11192–11203, 2019.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320, 2019.

Cullina, D., Bhagoji, A. N., and Mittal, P. Pac-learning in the presence of adversaries. In *Advances in Neural Information Processing Systems*, pp. 230–241, 2018.

Diochnos, D., Mahloujifar, S., and Mahmoody, M. Adversarial risk and robustness: General definitions and implications for the uniform distribution. In *Advances in Neural Information Processing Systems*, pp. 10359–10368, 2018.

Diochnos, D. I., Mahloujifar, S., and Mahmoody, M. Lower bounds for adversarially robust pac learning. In *ISAIM*, 2020.

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

Fawzi, A., Fawzi, H., and Fawzi, O. Adversarial vulnerability for any classifier. In *Advances in Neural Information Processing Systems*, pp. 1178–1187, 2018.

Garg, S., Jha, S., Mahloujifar, S., and Mohammad, M. Adversarially robust learning could leverage computational hardness. In *Algorithmic Learning Theory*, pp. 364–385, 2020.

Gilmer, J., Metz, L., Faghri, F., Schoenholz, S. S., Raghu, M., Wattenberg, M., and Goodfellow, I. Adversarial spheres. In *ICLR workshop*, 2018.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

Graybill, F. A. An introduction to linear statistical models. Technical report, 1961.

Gu, S. and Rigazio, L. Towards deep neural network architectures robust to adversarial examples. In *ICLR workshop*, 2015.

He, W., Wei, J., Chen, X., Carlini, N., and Song, D. Adversarial example defense: Ensembles of weak defenses are not strong. In *11th USENIX Workshop on Offensive Technologies (WOOT 17)*, 2017.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pp. 125–136, 2019.

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

Kos, J., Fischer, I., and Song, D. Adversarial examples for generative models. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 36–42. IEEE, 2018.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 656–672. IEEE, 2019.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Mahloujifar, S., Diochnos, D. I., and Mahmoody, M. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4536–4543, 2019.

Montasser, O., Hanneke, S., and Srebro, N. Vc classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, pp. 2512–2530, 2019.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Najafi, A., Maeda, S.-i., Koyama, M., and Miyato, T. Robustness to adversarial perturbations in learning from incomplete data. In *Advances in Neural Information Processing Systems*, pp. 5541–5551, 2019.

Nakkiran, P. Adversarial robustness may be at odds with simplicity. *arXiv preprint arXiv:1901.00532*, 2019.

Owhadi, H. and Scovel, C. Brittleness of bayesian inference and new selberg formulas. *Communications in Mathematical Sciences*, 14(1):83–145, 2016.

Owhadi, H. and Scovel, C. Qualitative robustness in bayesian inference. *ESAIM: Probability and Statistics*, 21:251–274, 2017.

Owhadi, H., Scovel, C., and Sullivan, T. On the brittleness of bayesian inference. *SIAM Review*, 57(4):566–582, 2015a.

Owhadi, H., Scovel, C., Sullivan, T., et al. Brittleness of bayesian inference under finite information in a continuous world. *Electronic Journal of Statistics*, 9(1):1–79, 2015b.

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

Papernot, N., McDaniel, P., Sinha, A., and Wellman, M. P. Sok: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 399–414. IEEE, 2018.

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. In *ICLR*, 2018a.

Raghunathan, A., Steinhardt, J., and Liang, P. S. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*, pp. 10877–10887, 2018b.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J. C., and Liang, P. Adversarial training can hurt generalization. In *ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena*, 2019.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J., and Liang, P. Understanding and mitigating the tradeoff between robustness and accuracy. *ICML*, 2020.

Rudin, W. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pp. 5014–5026, 2018.

Schulz, J. *The optimal Berry-Esseen constant in the binomial case*. Dissertation, Universität Trier, 2016.

Shafahi, A., Huang, W. R., Studer, C., Feizi, S., and Goldstein, T. Are adversarial examples inevitable? In *ICLR*, 2018.

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

Shevtsova, I. On the absolute constants in the berry-esseen-type inequalities. *Doklady Mathematics*, 89(3):378–381, 2014.

Stutz, D., Hein, M., and Schiele, B. Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6976–6987, 2019.

Stutz, D., Hein, M., and Schiele, B. Confidence-calibrated adversarial training: Generalizing to unseen attacks. In *ICML*, 2020.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *ICLR*, 2014.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *ICLR*, 2019.

Weng, L., Zhang, H., Chen, H., Song, Z., Hsieh, C.-J., Daniel, L., Boning, D., and Dhillon, I. Towards fast computation of certified robustness for relu networks. In *International Conference on Machine Learning*, pp. 5276–5285, 2018.

Wong, E. and Kolter, Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pp. 5283–5292, 2018.

Xu, H. and Mannor, S. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

Yin, D., Kannan, R., and Bartlett, P. Rademacher complexity for adversarially robust generalization. In *ICML*, pp. 7085–7094, 2019.

Zhang, H., Weng, T.-W., Chen, P.-Y., Hsieh, C.-J., and Daniel, L. Efficient neural network robustness certification with general activation functions. In *Advances in neural information processing systems*, pp. 4939–4948, 2018.

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, 2019.

Zhang, J., Xu, X., Han, B., Niu, G., Cui, L., Sugiyama, M., and Kankanhalli, M. Attacks which do not kill training make adversarial learning stronger. In *ICML*, 2020.