# Sample Complexity of Robust Linear Classification on Separated Data

**Robi Bhattacharjee** [1] **Somesh Jha** [2] **Kamalika Chaudhuri** [1]

## Abstract

We consider the sample complexity of learning with adversarial robustness. Most prior theoretical results for this problem have considered a setting where different classes in the data are close together or overlapping. We consider, in contrast, the well-separated case where there exists a classifier with perfect accuracy and robustness, and show that the sample complexity narrates an entirely different story. Specifically, for linear classifiers, we show a large class of well-separated distributions where the expected robust loss of any algorithm is at least $\Omega(\frac{d}{n})$, whereas the max margin algorithm has expected standard loss $O(\frac{1}{n})$. This shows a gap in the standard and robust losses that cannot be obtained via prior techniques. Additionally, we present an algorithm that, given an instance where the robustness radius is much smaller than the gap between the classes, gives a solution with expected robust loss is $O(\frac{1}{n})$. This shows that for very well-separated data, convergence rates of $O(\frac{1}{n})$ are achievable, which is not the case otherwise. Our results apply to robustness measured in any $\ell_p$ norm with $p > 1$ (including $p = \infty$).

## 1. Introduction

Motivated by the use of machine learning in safety-critical settings, adversarially robust classification has been of much recent interest. Formally, the problem is as follows. A learner is given training data drawn from an underlying distribution $D$, a hypothesis class $\mathcal{H}$, a robustness metric $d$, and a radius $r$. The learner's goal is to find a classifier $h \in \mathcal{H}$ which has the lowest robust loss at radius $r$. The robust loss of a classifier is the expected fraction of examples where either $f(x) \neq y$ or where there exists an $x'$ at distance $d(x, x') \leq r$ such that $f(x) \neq f(x')$. Robust classification thus aims to find a classifier that maximizes accuracy on examples that are distance $r$ or more from the decision boundary, where distances are measured according to the metric $d$.

In this work, we ask: how many samples are needed to learn a classifier with low robust loss when $\mathcal{H}$ is the class of linear classifiers, and $d$ is an $\ell_p$-metric? Prior work has provided both upper (Yin et al., 2019; Dan et al., 2020) as well as lower bounds (Schmidt et al., 2018; Dan et al., 2020) on the sample complexity of the problem. However, almost all look at settings where the data distribution itself is not separated – data from different classes overlap or are close together in space. In this case, the classifier that minimizes robust loss is quite different from the one that minimizes error, which often leads to strong sample complexity gaps. Many real tasks where robust solutions are desired however tend to involve well-separated data (Yang et al., 2020), and hence it is instructive to look at what happens in these cases.

With this motivation, we consider in this work robust classification of data that is linearly $r$-separable. Specifically, there exists a linear classifier which has zero robust loss at robustness radius $r$. This case is thus the analog of the realizable case for robust classification, and we consider both upper and lower bounds in this setting.

For lower bounds, prior work (Cullina et al., 2018) shows that both standard and robust linear classification have VC-dimension $O(d)$, and consequently have similar bounds on the expected loss in the worst case. However, these results do not apply to this setting since we are specifically considering well-separated data, which greatly restricts the set of possible worst-case distributions. For our lower bound, we provide a family of distributions that are linearly $r$-separable and where the maximum margin classifier, given $n$ independent samples, has error $O(1/n)$. In contrast, any algorithm for finding the minimum robust loss classifier has robust loss at least $\Omega(d/n)$, where $d$ is the data dimension. These bounds hold for all $\ell_p$-norms provided $p > 1$, including $p = 2$ and $p = \infty$. Unlike prior work, our bounds do not rely on the difference in loss between the solutions with optimal robust loss and error, and hence cannot be obtained by prior techniques. Instead, we introduce a new geometric construction that exploits the fact that learning a classifier with low robust loss when data is linearly $r$-separated requires

*Equal contribution [1]University of California, San Diego [2]University of Wisconsin-Madison. Correspondence to: Robi Bhattacharjee <rcbhatta@eng.ucsd.edu>.

*Proceedings of the $38^{th}$ International Conference on Machine Learning*, PMLR 139, 2021. 

seeing a certain number of samples close to the margin.

For upper bounds, prior work (Yin et al., 2019) provides a bound on the Rademacher complexity of adversarially robust learning, and show that it can be worse than the standard Rademacher complexity by a factor of $d^{1/q}$ for $\ell_p$-norm robustness where $1/p + 1/q = 1$. Thus, an interesting question is whether dimension-independent bounds, such as those for the accuracy under large margin classification, can be obtained for robust classification as well. Perhaps surprisingly, we show that when data is really well-separated, the answer is yes. Specifically, if the data distribution is linearly $r + \gamma$-separable, then there exists an algorithm that will find a classifier with robust loss $O(\Delta^2/\gamma^2 n)$ at radius $r$ where $\Delta$ is the diameter of the instance space. Observe that much like the usual sample complexity results on SVM and perceptron, this upper bound is independent of the data dimension and depends only on the excess margin (over $r$). This establishes that when data is really well-separated, finding robust linear classifiers does not require a very large number of samples.

While the main focus of this work is on linear classifiers, we also show how to generalize our upper bounds to Kernel Classification, where we find a similar dynamic with the loss being governed by the excess margin in the embedded kernel space. However, we defer a thorough investigation of robust kernel classification as an avenue for future work.

Our results imply that while adversarially robust classification may be more challenging than simply accurate classification when the classes overlap, the story is different when data is well-separated. Specifically, when data is linearly (exactly) $r$-separable, finding an $r$-separated solution to robust loss $\epsilon$ may require $\Omega(d/\epsilon)$ samples for some distribution families where finding an accurate solution is easier. Thus in this case, there is a gap between the sample complexities of robust and simply accurate solutions, and this is true regardless of the $\ell_p$ norm in which robustness is measured. In contrast, if data is even more separated – linearly $r + \gamma$-separable – then we can obtain a dimension-independent upper bound on the sample complexity, much like the sample complexity of SVMs and perceptron. Thus, how separable the data is matters for adversarially robust classification, and future works in the area should consider separability while discussing the sample complexity

### 1.1. Related Work

There is a large body of work (Carlini & Wagner, 2017; Liu et al., 2017; Papernot et al., 2017; 2016a; Szegedy et al., 2014; Hein & Andriushchenko, 2017; Katz et al., 2017; Papernot et al., 2016b; Raghunathan et al., 2018; Sinha et al., 2018) empirically studying adversarial examples primarily in the context of neural networks. Several works (Schmidt et al., 2018; Raghunathan et al., 2020; Tsipras et al., 2019) have empirically investigated trade-offs between robust and standard classification.

On the theoretical side, this phenomenon has been studied in both the parametric and non-parametric settings. On the parametric side, several works (Khim & Loh, 2018; Attias et al., 2019; Montasser et al., 2019; Yin et al., 2019; Ashtiani et al., 2020) have focused on finding distribution agnostic bounds of the sample complexity for robust classification. In (Montasser et al., 2019), Srebro et. al. showed through an example that the VC dimension of robust learning may be much larger than standard or accurate learning indicating that the sample complexity bounds may be higher. However, their example did not apply to linear classifiers.

(Diakonikolas et al., 2020) considers learning linear classifiers robustly, but is primarily focused on computational complexity as opposed to sample complexity.

In (Yin et al., 2019), Bartlett et. al. investigated the Rademacher complexity of robustly learning linear classifiers as well as neural networks. They showed that in both cases, the robust Rademacher complexity can be bounded in terms of the dimension of the input space – thus indicating a possible gap between standard and robust learning. However, as with the works considering VC dimension, this work is fundamentally focused on upper bounds – they do not show true lower bounds on data requirements.

Because of it's simplicity and elegance, the case where the data distribution is a mixture of Gaussians has been particularly well-studied. The first such work was (Schmidt et al., 2018), in which Schmidt et. al. showed an $\Omega(\sqrt{d})$ gap between the standard and robust sample complexity for a mixture of two Gaussians using the $\ell_\infty$ norm. This was subsequently expanded upon in (Bhagoji et al., 2019), (Dobriban et al., 2020) and (Dan et al., 2020). (Bhagoji et al., 2019) introduces a notion of "optimal transport," which they subsequently apply to the Gaussian case, deriving a closed form expression for the optimally robust linear classifier. Their results apply to any $\ell_p$ norm. (Dobriban et al., 2020) applies expands upon (Schmidt et al., 2018) by consider mixtures of three Gaussians in both the $\ell_2$ and $\ell_\infty$ norms. Finally, (Dan et al., 2020) fully generalizes the results of (Schmidt et al., 2018) providing tight upper and lower bounds on the standard and robust sample complexities of a mixture of two Gaussians, in any norm (including $\ell_p$ for $p \in [1, \infty]$). (Schmidt et al., 2018) and (Dan et al., 2020) bear the most relevance with our work, and we consequently carefully compare our results in section 3.1.

Another approach for lower and upper bounds on sample complexities for linear classifiers can be found in (Cullina et al., 2018), which examines the robust VC dimension of learning linear classifiers. They show that the VC dimension is $d+1$, just as it is in the standard case. This implies that the

bounds in the robust case match the bounds in the standard case and in particular shows a lower bound of $\Omega(d/n)$ on the expected loss of learning a robust linear classifier from $n$ samples.

While this result appears to match our lower bound, there is a crucial distinction between the bounds. Our bound implies that there exists some distribution with a large $\ell_2$ margin for which the expected robust loss must be $\Omega(d/n)$. On the other hand, standard results about learning linear classifiers on large margin data implies that the expected standard loss will be $O(1/n)$ (when running the max-margin algorithm). For this reason, our paper provides a case in the well-separated setting in which learning linear classifiers is provably more difficult (in terms of sample complexity) in the robust setting than in the standard setting. By contrast, (Cullina et al., 2018) does not show this. Their paper only implies (through standard VC constructions) the existence of *some* distribution that is difficult to learn, and the standard PAC bounds cannot ensure that such a distribution also has a large $\ell_2$ margin.

In the non-parametric setting, there are several works which contrast standard learning with robust learning. (Wang et al., 2018) considers the nearest neighbors algorithm, and shows how to adapt it for converging towards a robust classifier. In (Yang et al., 2019), Yang et. al. propose the *$r$-optimal classifier*, which is the robust analog of the Bayes optimal classifier. Through several examples they show that it is often a fundamentally different classifier - which can lead to different convergence behavior in the standard and robust settings. (Bhattacharjee & Chaudhuri, 2020) unified these approaches by specifying conditions under which non-parametric algorithms can be adapted to converge towards the $r$-optimal classifier, thus introducing $r$-consistency, the robust analog of consistency.

## 2. Preliminaries

We consider binary classification over $\mathbb{R}^d \times \{\pm 1\}$. Our metric of choice is the $\ell_p$ norm, where $p > 1$ (including $p = \infty$) is arbitrary. For $x \in \mathbb{R}^d$, we will use $||x||_p$ to denote the $\ell_p$ norm of $x$, and consequently will use $||x - y||_p$ to denote the $\ell_p$ distance between $x$ and $y$. We will also let $\ell_q$ denote the dual norm to $\ell_p$ - that is, $\frac{1}{q} + \frac{1}{p} = 1$.

We use $B_p(x, r)$ to denote the closed $\ell_p$ ball with center $x$ and radius $r$. For any $S \subset \mathbb{R}^d$, we let $diam_p(S)$ denote its diameter: that is, $diam_p(S) = \sup_{x,y \in S} ||x - y||_p$.

### 2.1. Standard and Robust Loss

In classical statistical learning, the goal is to learn an accurate classifier, which is defined as follows:

**Definition 1.** *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d \times \{\pm 1\}$, and let $f \in \{\pm 1\}^{\mathbb{R}^d}$ be a classifier. Then the **standard loss** of $f$ over $\mathcal{D}$, denoted $\mathcal{L}(f, \mathcal{D})$, is the fraction of examples $(x, y) \sim \mathcal{D}$ for which $f$ is not accurate. Thus*

$$\mathcal{L}(f, \mathcal{D}) = P_{(x,y)\sim\mathcal{D}}[f(x) \neq y].$$

Next, we define robustness, and the corresponding robust loss.

**Definition 2.** *A classifier $f \in \{\pm 1\}^{\mathbb{R}^d}$ is said to be **robust** at $x$ with radius $r$ if $f(x) = f(x')$ for all $x' \in B_p(x, r)$.*

**Definition 3.** *The **robust loss** of $f$ over $\mathcal{D}$, denoted $\mathcal{L}_r(f, \mathcal{D})$, is the fraction of examples $(x, y) \sim \mathcal{D}$ for which $f$ is either inaccurate at $(x, y)$, or $f$ is not robust at $(x, y)$ with radius $r$. Observe that this occurs if and only if there is some $x' \in B_p(x, r)$ such that $f(x') \neq y$. Thus*

$$\mathcal{L}_r(f, \mathcal{D}) = P_{(x,y)\sim\mathcal{D}}[\exists x' \in B_p(x, r) \text{ s.t. } f(x') \neq y].$$

### 2.2. Expected Loss and Sample Complexity

The most common way to characterize the performance of a learning algorithm is through an $(\epsilon, \delta)$ guarantee, which computes $\epsilon_n, \delta_n$ such that an algorithm trained over $n$ samples has loss at most $\epsilon_n$ with probability at least $1 - \delta_n$.

In this work, we use the simpler notion of *expected loss*, which is defined as follows:

**Definition 4.** *Let $A$ be a learning algorithm and let $\mathcal{D}$ be a distribution over $\mathbb{R}^d \times \{\pm 1\}$. For any $S \sim \mathcal{D}^n$, we let $A_S$ denote the classifier learned by $A$ from training data $S$. Then the **expected standard loss** of $A$ with respect to $\mathcal{D}$, denoted $EL^n(A, \mathcal{D})$ where $n$ is the number of training samples, is defined as*

$$E\mathcal{L}^n(A, \mathcal{D}) = \mathbb{E}_{S\sim\mathcal{D}^n}\mathcal{L}(A_S, \mathcal{D}).$$

*Similarly, we define the **expected robust loss** of $A$ with respect to $\mathcal{D}$ as*

$$E\mathcal{L}_r^n(A, \mathcal{D}) = \mathbb{E}_{S\sim\mathcal{D}^n}\mathcal{L}_r(A_S, \mathcal{D}).$$

Our main motivation for using this criteria is simplicity. Our primary goal is to compare and contrast the performances of algorithms in the standard and robust cases, and this contrast clearest when the performances are summarized as a single number (namely the expected loss) rather than an $(\epsilon, \delta)$ pair.

Next, we address the notion of sample complexity. As above, sample complexity is typically defined as the minimum number of samples needed to guarantee $(\epsilon, \delta)$ performance. In this work, we will instead define it solely with respect to $\epsilon$, the expected loss.

**Definition 5.** *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d \times \{\pm 1\}$ and $A$ be a learning algorithm. Then the **standard sample complexity** of $A$ with respect to $\mathcal{D}$, denoted $m^\epsilon(A, \mathcal{D})$, is the*

*minimum number of training samples needed such that $A$ has expected standard loss at most $\epsilon$. Formally,*

$$m^{\epsilon}(A, \mathcal{D}) = \min(\{n : E\mathcal{L}^n(A, D) \leq \epsilon\}).$$

*Similarly, we can define the **robust sample complexity** as*

$$m_r^{\epsilon}(A, \mathcal{D}) = \min(\{n : E\mathcal{L}^n(A, D) \leq \epsilon\}).$$

### 2.3. Linear classifiers

In this work, we consider linear classifiers, formally defined as follows:

**Definition 6.** *Let $w \in \mathbb{R}^d$ be a vector. Then the **linear classifier** with parameters $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ over $\mathbb{R}^d \times \pm 1$, denoted $f_{w,b}$, is defined as ,*

$$f_{w,b}(x) = \begin{cases} +1 & \langle w, x \rangle \geq b \\ -1 & \langle w, x \rangle < b \end{cases}.$$

Learning linear classifiers is well understood in the standard classification setting. We now consider the linearly *separable* case, in which some linear classifier has perfect accuracy. We will later define linear $r$-separability as the robust analog of separability.

**Definition 7.** *A distribution $\mathcal{D}$ over $\mathbb{R}^d \times Y$ is **linearly separable** if its support can be partitioned into sets $S^+$ and $S^-$ such that:*

*1. $S^+$ and $S^-$ correspond to the positively and negatively labeled subsets of $\mathbb{R}^d$. In particular, $P_{(x,y)\sim\mathcal{D}}[x \in S^y] = 1$.*

*2. There exists a linear classifier, $f_{w,b}$, that has perfect accuracy. That is, $\mathcal{L}(f_{w,b}, \mathcal{D}) = 0$.*

The standard sample complexity for linearly separable distributions can be characterized through their margin, which is defined as follows.

**Definition 8.** *Let $\mathcal{D}$ be a linearly separable distribution over $\mathbb{R}^d \times \{\pm 1\}$. Let $S^+$ and $S^-$ be as above. Then $\mathcal{D}$ has **margin** $\gamma$ if $\gamma$ is the largest real number such that there exists a linear classifier $f_{w,b}$ with the following properties:*

*1. $f_{w,b}$ has perfect accuracy. That is, $\mathcal{L}(f_{w,b}, \mathcal{D}) = 0$.*

*2. Let $H_{w,b} = \{x : \langle x, w \rangle = b\}$ denote the decision boundary of $f_{w,b}$. Then for all $x \in (S^+ \cup S^-)$, $x$ has $\ell_2$ distance at least $\gamma$ from $H_{w,b}$. That is,*

$$\inf_{x \in S^+ \cup S^-, z \in H_{w,b}} ||x - z||_2 \geq \gamma.$$

*We let $\gamma(\mathcal{D})$ denote the margin of $\mathcal{D}$.*

Observe that although we use a general norm, $\ell_p$, to measure robustness, the margin is always measured in $\ell_2$. This is because the $\ell_2$ norm plays a fundamental role in bounding the number of samples needed to learn a linear classifier.

The basic idea is that when the $\ell_2$ margin is large relative to the $\ell_2$ diameter of the distribution, the max margin algorithm requires fewer samples needed to learn a linear classifier. In particular, the ratio between the $\ell_2$ margin and the $\ell_2$ diameter fully characterizes the standard sample complexity of the max margin algorithm. To further simplify our notation, we define this ratio as the aspect ratio.

**Definition 9.** *Let $\mathcal{D}$ be a linearly separable distribution over $\mathbb{R}^d \times \{\pm 1\}$. Then the **aspect ratio** of $\mathcal{D}$, $\rho(\mathcal{D})$ is defined as,*

$$\rho(\mathcal{D}) = \frac{diam_2(S^+ \cup S^-)}{\gamma(\mathcal{D})},$$

*where $diam_2(S^+ \cup S^-)$ denotes its diameter in the $\ell_2$ norm.*

We now have the following well-known result, which characterizes the expected standard loss with the aspect ratio.

**Theorem 10.** (Chapter 10 in (Vapnik, 1998)) *Let $M$ denote the hard margin SVM algorithm. If $\mathcal{D}$ is a distribution with aspect ratio $\rho = \rho(\mathcal{D})$, then for any $n > 0$ we have $\mathbb{E}_{S\sim\mathcal{D}^n}\mathcal{L}(M_S, \mathcal{D}) \leq O(\frac{\rho^2}{n})$, where $M_S$ denotes the classifier learned by $M$ from training data $S$.*

We can also express this result in terms of standard sample complexity.

**Corollary 11.** *Let $M$ denote the hard margin SVM algorithm. If $\mathcal{D}$ is a distribution with aspect ratio $\rho = \rho(\mathcal{D})$, then for any $\epsilon > 0$ we have $m^{\epsilon}(M_S, \mathcal{D}) \leq O(\frac{\rho^2}{\epsilon})$, where $M_S$ denotes the classifier learned by $M$ from training data $S$.*

Theorem 10 and Corollary 11 will serve as a benchmark for comparison with the robust sample complexity.

### 2.4. Linear $r$-separability

Finally, we introduce linear $r$-separability, which is the key characteristic of distributions considered in this paper. This can be thought of as the robust analog of linear separability.

**Definition 12.** *For any $r > 0$, a distribution $\mathcal{D}$ over $\mathbb{R}^d \times \{\pm 1\}$ is **linearly $r$-separable** if there exists a linear classifier $f_{w,b}$ such that $\mathcal{L}_r(f_{w,b}, \mathcal{D}) = 0$.*

This definition is the fundamental property considered in this paper. Our goal is to understand the sample complexity required for learning robust linear classifiers on linearly $r$-separable distributions, and compare it with the standard sample complexity given in Theorem 10.

## 3. Lower Bounds

In this section, we consider $r$-separated distributions whose aspect ratio is constant. By Theorem 10, the standard sample

complexity for learning them is independent of $d$. We will show that in contrast, the robust sample complexity has a linear dependence on $d$, and consequently establish a substantial gap between the standard and robust cases.

We begin by defining the family of such distributions.

**Definition 13.** *For any $\rho, r$, the set $\mathcal{F}_{r,\rho}$ is defined as the set of all distributions $\mathcal{D}$ over $\mathbb{R}^d \times \{\pm 1\}$ such that $\mathcal{D}$ is $r$-separated and has aspect ratio at most $\rho$.*

We now state our main result.

**Theorem 14.** *Let $r > 0$ and $\rho > 20$. Then the following hold.*

1. *For every learning algorithm $A$, and any $n > 0$, there exists $\mathcal{D} \in \mathcal{F}_{r,\rho}$ such that the expected robust loss when $A$ is trained on a sample of size $n$ from $\mathcal{D}$ is at least $\Omega(\frac{d}{n})$. Formally, there exists a constant $c > 0$ such that $\mathbb{E}_{S\sim\mathcal{D}^n}[\mathcal{L}_r(A_S, \mathcal{D})] \geq \frac{cd}{n}$.*
2. *In contrast, by Theorem 10, for any $\mathcal{D} \in \mathcal{F}_{r,D}$, the max margin algorithm has expected standard loss $O(\frac{\rho^2}{n})$, when trained on a sample of size $n$ from $\mathcal{D}$. Formally, there exists a constant $c' > 0$ such that $\mathbb{E}_{S\sim\mathcal{D}^n}[\mathcal{L}(A_S, \mathcal{D})] \leq \frac{c'\rho^2}{n}$.*

The condition $\rho > 20$ is required to rule out degenerate cases. This is because for small values of $\rho$, the $\ell_2$ diameter of $\mathcal{D}$ is not much larger than the $\ell_2$ margin of $\mathcal{D}$. This forces $\mathcal{D}$ to be mostly clustered around a line which leads to more complicated behavior.

Observe that when $\rho$ is a constant independent of $d$, the expected standard loss is $O(\frac{1}{n})$ while the expected robust loss is $\Omega(\frac{d}{n})$. Thus, the ratio between the expected robust loss and the expected standard loss is $\Omega(d)$, leading to a dimensional dependent gap between the robust and standard cases.

We also note that these bounds hold regardless of which $\ell_p$ ($p \in (1, \infty]$) norm is being used. This is because our construction of $\mathcal{D} \in \mathcal{F}_{r,\rho}$ for which the lower bound holds is given in terms of the norm $p$. More generally, the family $\mathcal{F}_{r,\rho}$ is implicitly defined with respect to $p$.

Furthermore, our lower bound differs from the lower bound of $\Omega(\frac{d}{n})$ shown in prior work (Cullina et al., 2018) because it specifically holds for $\mathcal{F}_{r,\rho}$, a linearly $r$-separated family of distributions with constant aspect ratio. Thus, while (Cullina et al., 2018) has shown the existence of distributions satisfying the first condition of Theorem 14, our result is the first to exhibit a distribution satisfying both conditions.

Finally, we note that Theorem 14 can also be expressed in terms of sample complexities. We include this in the following corollary.

**Corollary 15.** *Let $r > 0$ and $\rho > 20$. Then the following hold.*

1. *For every learning algorithm $A$, and any $\epsilon > 0$, there exists $\mathcal{D} \in \mathcal{F}_{r,\rho}$ such that the robust sample complexity of $A$ with respect to $\mathcal{D}$ is at least $\Omega(\frac{d}{\epsilon})$. Formally, there exists a constant $c > 0$ such that $m_r^\epsilon(A, \mathcal{D}) \geq \frac{cd}{\epsilon}$.*
2. *In contrast, by Theorem 10, for any $\mathcal{D} \in \mathcal{F}_{r,D}$, the max margin algorithm has standard sample complexity $O(\frac{\rho^2}{\epsilon})$. Formally, there exists a constant $c' > 0$ such that $m^\epsilon(A, \mathcal{D}) \leq \frac{c'\rho^2}{\epsilon}$.*

### 3.1. Comparison with (Dan et al., 2020) and (Schmidt et al., 2018)

The first work to provide a robust sample complexity lower bound that applied to linear classifiers is (Schmidt et al., 2018); they showed a gap of $\Omega(\sqrt{d})$ between the robust and accuracy loss for a specific mixture of two Gaussians. This was later generalized to mixtures of any two Gaussians by (Dan et al., 2020), who also established more general lower bounds for any $\ell_p$ norm. Since (Dan et al., 2020) is a strict generalization of (Schmidt et al., 2018), we next explain how our lower bounds differ from (Dan et al., 2020), and why their techniques do not lead to our results. We begin by summarizing their results.

**Summary of (Dan et al., 2020)** (Dan et al., 2020) considers data distributions $\mathcal{D}$ that are parametrized by $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$, $\Sigma \succcurlyeq 0$. $\mathcal{D}_{\mu,\Sigma}$ is the mixture of two Gaussians, $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(-\mu, \Sigma)$, with equal mass, where instances drawn from $\mathcal{N}(\mu, \Sigma)$ are labeled as $+$, and instances drawn from $\mathcal{N}(-\mu, \Sigma)$ are labeled as $-$. They consider robustness measured in any normed metric in $\mathbb{R}^d$, including the $\ell_p$ norm for $p \in (1, \infty]$. Although their bounds apply to any classifier, this effectively deals with linear classifiers since it can be shown that the optimally robust and accurate classifiers are both linear.

For any distribution $\mathcal{D}_{\mu,\Sigma}$, let $L_{rob}$ denote the optimal robust loss of any classifier on $\mathcal{D}_{\mu,\Sigma}$, and let $L_{std}$ denote the optimal standard loss. Then the bounds shown in (Dan et al., 2020) can restated as follows (a detailed derivation from (Dan et al., 2020) appears in Appendix A).

**Theorem 16.** *(Dan et al., 2020)*

1. *For any learning algorithm $A$ and any $n > 0$, there exists some mixture of Gaussians, $\mathcal{D}_{\mu,\Sigma}$ such that the expected excess robust loss is at least $\Omega(L_{rob}\frac{d}{n})$, when $A$ is trained on a sample of size $n$ from $\mathcal{D}$.*
2. *For any distribution $\mathcal{D}_{\mu,\Sigma}$, it is possible to learn a classifier with expected excess standard loss at most $O(L_{std}\frac{d}{n})$.*

3. *By (1.) and (2.), the ratio between the expected excess loss and expected excess standard loss can be expressed as* $ratio \geq \Omega(\frac{L_{rob}}{L_{std}})$.

Observe that their bounds are given through *excess* losses, which is the amount by which the loss exceeds to the optimal loss. This is necessary because in their setting, the optimal classifiers do not have 0 loss.

**Comparison with our bounds** Recall that in our work, we are concerned with the *linearly $r$-separated case*, which occurs precisely when the optimal robust and standard losses both equal 0. However, from Theorem 16, we see that although (Dan et al., 2020) proves a gap between standard and robust sample complexity, this gap is predicated on distributions for which the optimal robust loss, $L_{rob}$ and optimal standard loss, $L_{std}$ differ. Furthermore, in the case where they obtain a gap of $\Omega(d)$, we see that this requires $\frac{L_{rob}}{L_{std}} = \Omega(d)$ which is a substantial difference. By contrast, our results characterize a gap exclusively in the case that this does not occur.

Finally, in the limiting case where the Gaussians they consider are sufficiently far apart, their data will begin to appear linearly $r$-separated, meaning both $L_{rob}$ and $L_{std}$ are close to 0. However, even in this case, it can be shown that the ratio $\frac{L_{rob}}{L_{std}}$ diverges towards infinity, meaning that their lower bound characterizes a very different dynamic from ours. Precise details on this comparison can be found in appendix A.

### 3.2. Intuition behind Theorem 14

The proof idea for Theorem 14 can be summarized with a simple example (Figure 1). In this example, we seek to learn a linear classifier for a linearly $r$-separated distribution in $\mathbb{R}^2$. The key idea is to contrast the necessary conditions for learning a robust classifier, and the necessary conditions for learning an accurate classifier.

Observe that the distribution is *precisely* linearly $r$-separated, that is, it is not possible to achieve robustness for radii larger than $r$. Because of this, there is a unique linear classifier $f_{rob}$ that has perfect robustness. In order to learn this classifier, we must see examples from $S^+ \cup S^-$ that are close to the "boundary" of $S^+ \cup S^-$. In our figure, this consists of points that are close to the dotted blue and red lines. Moreover, it can be shown that the number of such examples we must see is related to $d$, the dimension.

By contrast, any classifier that separates $S^+$ from $S^-$ has perfect accuracy (take for example $f_{std}$ shown in the figure). It is possible to exploit this by using margin based algorithms for learning linear classifiers. In particular, we no longer need to see points that are extremely close to the boundary of $S^+ \cup S^-$.

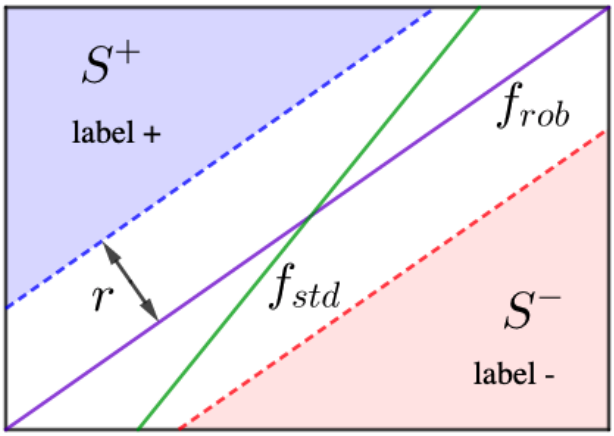


*Figure 1.* An example of a linearly $r$-separated distribution, with positively and negatively labeled examples in $S^+$ and $S^-$ respectively. The optimally robust classifier, $f_{rob}$ is shown in purple, while the (not necessarily unique) optimally accurate classifier, $f_{std}$, is shown in green.

**General Hypothesis Classes:** We now briefly consider how to extend our methods to other hypothesis classes. For any hypothesis class $\mathcal{H}$ and distribution $\mathcal{D}$ let

$$\mathcal{H}_{\mathcal{D},\alpha} = \{h : h \in \mathcal{H}, \mathcal{L}(h, \mathcal{D}) \leq \alpha\}$$

and let

$$\mathcal{H}^r_{\mathcal{D},\alpha} = \{h : h \in \mathcal{H}, \mathcal{L}_r(h, \mathcal{D}) \leq \alpha\}.$$

$\mathcal{H}_{\mathcal{D},\alpha}$ can be thought of as the set of accurate classifiers while $\mathcal{H}^r_{\mathcal{D},\alpha}$ can be thought of as the set of astute classifiers. By their definitions, it is clear that $\mathcal{H}^r_{\mathcal{D},\alpha} \subseteq \mathcal{H}_{\mathcal{D},\alpha}$. However, in the case when $\mathcal{H}$ is the set of linear classifiers, we see that for small $\alpha$, $\mathcal{H}^r_{\mathcal{D},\alpha}$ is a much "smaller" set than $\mathcal{H}_{\mathcal{D},\alpha}$. By exploiting the geometric structure inherent to $\mathcal{H}$, we can much more efficiently search for some $h \in \mathcal{H}_{\mathcal{D},\alpha}$ than we can in $\mathcal{H}^r_{\mathcal{D},\alpha}$. This dynamic is the crux of our lower bound: as we essentially show that there are far more critical points (i.e. points near the decision boundary) that we must see for learning $\mathcal{H}^r_{\mathcal{D},\alpha}$ that aren't required for $\mathcal{H}_{\mathcal{D},\alpha}$.

Thus, for our methods to extend to an arbitrary hypothesis class, we would require a similar dynamic. We need two properties to hold: (1) $\mathcal{H}^r_{\mathcal{D},\alpha}$ must be a very strict subset of $\mathcal{H}_{\mathcal{D},\alpha}$ for sufficiently small alpha. (2) We must have some kind of exploitable geometric structure about $\mathcal{H}$ which allows us to exploit this gap. For the case of linear classifiers, this was the $\ell_2$ measured aspect ratio, $\gamma(\mathcal{D})$.

**Algorithm 1** Adversarial-Perceptron

**Input**: $S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \sim \mathcal{D}^n$,
$w \leftarrow 0$
**for** $i = 1 \ldots n$ **do**
  $z = \arg\min_{||z - x_i||_p \leq r} y_i \langle w, z \rangle$ {finds adv. ex.}
  **if** $\langle w, y_i z \rangle \leq 0$ {checks label} **then**
    $w \leftarrow w + y_i z$ {perceptron update}
  **end if**
**end for**
return $f_{w,0}$

*Figure 2.* An algorithm combining adversarial training with the perceptron algorithm. For each $(x_i, y_i)$, we first attack it, to get $z$. If $z$ is labeled incorrectly, we do a perceptron update using $z$.

**Kernel Classifiers:** A natural choice of a more general hypothesis class would be Kernel Classifiers, which are linear classifiers that operate in an embedded space, $H$. The main difficulty in expanding our lower bound to this more general setting comes from the behavior near the margin: the effects of the robustness radius in the embedded space are considerably less behaved than they are in the standard linear case. Nevertheless, we leave this as an important avenue for future work.

## 4. Upper Bounds

In the previous section, we showed that for any algorithm, there is some distribution $\mathcal{D} \in \mathcal{F}_{r,\rho}$ that is difficult (i.e. requires high sample complexity) to learn robustly. A natural follow-up question is: what about distributions for which the margin, $\gamma$ is very large compared to $r$.

Observe that in Figure 1 the robustness radius $r$ is very close to the margin. In particular, we can find adversarial examples from $S^+$ and $S^-$ that are very close to the decision boundary $f_{rob}$. By contrast, if $\gamma >> r$, then this no longer holds which suggests that better robust sample complexities might be possible.

In this section, we will describe a subset of $\mathcal{F}_{r,\rho}$ that can be learned with expected loss $O(\frac{1}{n})$, thus matching the standard sample complexity up to a constant factor. To do so, we will introduce a novel concept: the *robust margin*. The basic intuition is that distributions for which the margin greatly exceeds the robustness radius are precisely distributions with a large robust margin. We use the following notation.

Observe that if $\mathcal{D}$ is a linearly $r$-separated distribution, then $\mathcal{D}$ must also be linearly separable. As earlier, let $S^+, S^- \subset \mathbb{R}^d$ denote the positively and negatively labeled examples from $\mathcal{D}$. We now define

$$S_r^+ = \cup_{s \in S^+} B_p(s, r) \text{ and } S_r^- = \cup_{s \in S^-} B_p(s, r). \quad (1)$$

It follows that the decision boundary of any linear classifier with perfect robustness over $\mathcal{D}$ must separate $S_r^+$ and $S_r^-$. We now define the robust margin as a measurement of this separation.

**Definition 17.** *Let $\mathcal{D}$ be a linearly $r$-separable distribution over $\mathbb{R}^d \times \{\pm 1\}$. Let $S_r^+$ and $S_r^-$ be as above. Then $\mathcal{D}$ has **robust margin** $\gamma_r$ if $\gamma_r$ is the largest real number such that there exists a linear classifier $f_{w,b}$ with the following properties:*

*1. $f_{w,b}$ has perfect astuteness. That is, $\mathcal{L}_r(f_{w,b}, \mathcal{D}) = 0$.*

*2. Let $H_{w,b} = \{x : \langle x, w \rangle = b\}$ denote the decision boundary of $f_{w,b}$. Then for all $x \in (S_r^+ \cup S_r^-)$, $x$ has $\ell_2$ distance at least $\gamma$ from $H_{w,b}$. That is,*

$$\inf_{x \in S_r^+ \cup S_r^-} \inf_{z \in H_{w,b}} ||x - z||_2 \geq \gamma.$$

*We let $\gamma_r(\mathcal{D})$ denote the margin of $\mathcal{D}$, and say that such a distribution is $r, \gamma_r$-separated.*

It is crucial to note that although adversarial perturbations are measured in $\ell_p$, the robust margin is measured in $\ell_2$. This is because while the metric $\ell_p$ plays a role in constructing $B(x, r)$, it can be completely disregarded once the sets $S_r^+$ and $S_r^-$ are considered, as any hyperplane separating $S_r^+$ and $S_r^-$ will have perfect robustness.

We now define the robust aspect ratio, which is the robust analog of standard aspect ratio.

**Definition 18.** *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d \times \{\pm 1\}$. Then the **robust aspect ratio** of $\mathcal{D}$, $\rho_r(\mathcal{D})$ is defined as*

$$\rho_r(\mathcal{D}) = \frac{diam_2(S_r^+ \cup S_r^-)}{\gamma_r(\mathcal{D})},$$

*where as before, $diam_2(S_r^+ \cup S_r^-)$ denotes its diameter in the $\ell_2$ norm.*

We will now show that just as the aspect ratio, $\rho(\mathcal{D})$, characterized the sample complexity for standard classification, the robust aspect ratio, $\rho_r(\mathcal{D})$ will characterize the sample complexity for robust learning. To do so, we present a perceptron-inspired algorithm (Algorithm 1) for learning a robust classifier on $r$-separated data with robust aspect ratio $\rho_r$.

The basic idea behind Algorithm 1 is to combine the standard perceptron algorithm with adversarial training. In particular, we iterate through the training set and do the following on each point (refer to Algorithm 1 for precise details).

1. Find an adversarial example $(z, y_i)$ by attacking our classifier, $f_{w,0}$, at $(x_i, y_i)$ (line 4). This is a straightforward convex optimization problem for linear classifiers.

2. If $f_{w,0}(z) \neq y_i$, we update our weight vector with $(z, y_i)$ by using the standard perceptron update (lines 5-6).

We have the following upper bound on the expected robust loss of our algorithm.

**Theorem 19.** *Let $\mathcal{D}$ be a distribution with robust aspect ratio $\rho_r(\mathcal{D})$. Then for any $n > 0$, we have*

$$\mathbb{E}_{S\sim\mathcal{D}^n}[\mathcal{L}_r(A_S, \mathcal{D})] \leq O(\frac{\rho_r(\mathcal{D})^2}{n}),$$

*where $A_S$ denotes the classifier learned by Algorithm 1 from training data $S$.*

Observe that this expected loss is still larger than the expected standard loss in Theorem 10 as $\rho_r(\mathcal{D}) > \rho(\mathcal{D})$ for any $\mathcal{D}$. We also note that this result is not contradictory with our lower bound; there exist distributions $\mathcal{D} \in \mathcal{F}_{r,\rho}$ such that $\gamma_r(\mathcal{D}) = 0$, and these are precisely the distributions for which our lower bounds hold.

### 4.1. Generalization to Kernel Classifiers

Algorithm 1 can be thought of as the robust analog to the perceptron algorithm. We now generalize this algorithm to obtain a robust variant of the *kernel perceptron algorithm*. We first briefly review kernel classifiers. A detailed explanation of our generalized algorithm along with requisite background material can be found in Appendix D

**Definition 20.** *Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a kernel similarity function, $T = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset \mathbb{R}^d \times \{\pm 1\}$ be a set of labeled points, and $\alpha \in \mathbb{R}^m$ be a vector of $m$ real numbers. Then the **kernel classifier** with similarity function $K$, parameters $T, \alpha$, and denoted by $f^{\alpha}_{T,K}$ is defined as*

$$f^K_{T,\alpha}(x) = \begin{cases} +1 & \sum_1^m \alpha_i y_i K(x_i, x) \geq 0 \\ -1 & \sum_1^m \alpha_i y_i K(x_i, x) < 0 \end{cases}.$$

Conceptually, kernel classifiers are linear classifiers operating in embedded space. With each kernel similarity function $K$, there is a map $\phi : \mathbb{R}^d \to H$ (where $H$ is some Hilbert space) such that $K(x, x') = \langle \phi(x), \phi(x') \rangle$. Thus we can think of kernel classifiers as having a linear decision boundary in $H$.

We now present an analog of Algorithm 1 that we call the Adversarial Kernel-Perceptron. The essence of this algorithm has not changed. For each $(x_t, y_t)$ in our training set, we do the following.

1. Find an adversarial example $(z, y_i)$ by attacking our classifier, $f^K_{T,\alpha}$, at $(x_i, y_i)$ (line 4).

**Algorithm 2** Adversarial-Kernel-Perceptron

1: **Input**: $S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \sim \mathcal{D}^n$, Similarity function, $K$
2: $T \leftarrow \emptyset$, $\alpha \leftarrow 0$
3: **for** $i = 1 \ldots n$ **do**
4: $\quad z = \arg\min_{||z-x||_p \leq r} y_i f^K_{T,\alpha}(z)$ {finds adv. ex.}
5: $\quad$ **if** $f^k_{T,\alpha}(z) \leq 0$ {checks label} **then**
6: $\quad\quad T = T \cup \{(z, y_i)\}$ {kern. percep. update}
7: $\quad\quad \alpha = (1, \ldots, 1)_{|T|}$
8: $\quad$ **end if**
9: **end for**
10: return $f^K_{T,\alpha}$

*Figure 3.* A kernel version of Algorithm 1. We replace the perceptron update step with a kernel-perceptron update step.

2. If $f^K_{T,\alpha}(z) \neq y_i$, we update our weight vector with $(z, y_i)$ by appending $(z, y_i)$ to $T$ (lines 5-6). This corresponds to a kernel-perceptron update that uses $(z, y_i)$ instead of $(x_i, y_i)$.

One challenging aspect of this algorithm is minimizing $f^k_{T,\alpha}(z)$. For linear classifiers, this has a closed form solution that utilizes the dual norm. For arbitrary Kernel classifiers, this is a somewhat more challenging problem. However, we note that this can be solved using standard optimization techniques, and in some cases (when $K$ is particularly simple), it can be solved with basic gradient descent.

Finally, we show that this Algorithm has similar performance to the linear case. Instead of using the robust aspect ratio, $\rho_r(\mathcal{D})$, to bound the performance, we will require the **robust $K$-aspect ratio**, which is the kernel analog of this quantity. It can be thought of as the robust aspect ratio in the embedded space $H$. Details about this quantity (along with the proof of the theorem) can be found in Appendix D.

**Theorem 21.** *Let $\mathcal{D}$ be a distribution with robust $K$-aspect ratio $\rho^K_r(\mathcal{D})$. Then for any $n > 0$, we have*

$$\mathbb{E}_{S\sim\mathcal{D}^n}[\mathcal{L}_r(A_S, \mathcal{D})] \leq O(\frac{\rho^K_r(\mathcal{D})^2}{n}),$$

*where $A_S$ denotes the classifier learned by Algorithm 2 from training data $S$.*

This result indicates that for small values of $\rho^k_r(\mathcal{D})$, we can achieve a very good robust sample complexity for kernel classifiers. However, as the size of the perturbations approach this margin, this quantity goes to infinity. This phenomenon mirrors the linearly separable case, and suggests that a similar overall dynamic holds for kernel classification. We leave finding a full generalization (including our lower bound) for a direction in future work.

## References


Ashtiani, H., Pathak, V., and Urner, R. Black-box certification and learning under adversarial perturbations. *CoRR*, abs/2006.16520, 2020. URL https://arxiv.org/abs/2006.16520.

Attias, I., Kontorovich, A., and Mansour, Y. Improved generalization bounds for robust learning. In Garivier, A. and Kale, S. (eds.), *Algorithmic Learning Theory, ALT 2019, 22-24 March 2019, Chicago, Illinois, USA*, volume 98 of *Proceedings of Machine Learning Research*, pp. 162–183. PMLR, 2019.

Bhagoji, A. N., Cullina, D., and Mittal, P. Lower bounds on adversarial robustness from optimal transport. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 7496–7508, 2019.

Bhattacharjee, R. and Chaudhuri, K. When are non-parametric methods robust? *CoRR*, abs/2003.06121, 2020. URL https://arxiv.org/abs/2003.06121.

Carlini, N. and Wagner, D. A. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pp. 39–57, 2017.

Cullina, D., Bhagoji, A. N., and Mittal, P. Pac-learning in the presence of adversaries. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 230–241. Curran Associates, Inc., 2018.

Dan, C., Wei, Y., and Ravikumar, P. Sharp statistical guarantees for adversarially robust gaussian classification. *CoRR*, abs/2006.16384, 2020. URL https://arxiv.org/abs/2006.16384.

Diakonikolas, I., Kane, D. M., and Manurangsi, P. The complexity of adversarially robust proper learning of halfspaces with agnostic noise. *CoRR*, abs/2007.15220, 2020. URL https://arxiv.org/abs/2007.15220.

Dobriban, E., Hassani, H., Hong, D., and Robey, A. Provable tradeoffs in adversarially robust classification. *CoRR*, abs/2006.05161, 2020. URL https://arxiv.org/abs/2006.05161.

Freund, Y. and Schapire, R. E. Large margin classification using the perceptron algorithm. *Mach. Learn.*, 37(3): 277–296, 1999.

Hein, M. and Andriushchenko, M. Formal guarantees on the robustness of a classifier against adversarial manipulation. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 2266–2276. Curran Associates, Inc., 2017.

Katz, G., Barrett, C. W., Dill, D. L., Julian, K., and Kochenderfer, M. J. Towards proving the adversarial robustness of deep neural networks. In *Proceedings First Workshop on Formal Verification of Autonomous Vehicles, FVAV@iFM 2017, Turin, Italy, 19th September 2017.*, pp. 19–26, 2017.

Khim, J. and Loh, P. Adversarial risk bounds for binary classification via function transformation. *CoRR*, abs/1810.09519, 2018. URL http://arxiv.org/abs/1810.09519.

Liu, Y., Chen, X., Liu, C., and Song, D. Delving into transferable adversarial examples and black-box attacks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Montasser, O., Hanneke, S., and Srebro, N. VC classes are adversarially robustly learnable, but only improperly. In Beygelzimer, A. and Hsu, D. (eds.), *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2512–2530. PMLR, 2019.

Papernot, N., McDaniel, P. D., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21-24, 2016*, pp. 372–387, 2016a.

Papernot, N., McDaniel, P. D., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, pp. 582–597, 2016b.

Papernot, N., McDaniel, P. D., Goodfellow, I. J., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against deep learning systems using adversarial examples. *ASIACCS*, 2017.

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J. C., and Liang, P. Understanding and mitigating the

tradeoff between robustness and accuracy. *CoRR*, abs/2002.10716, 2020. URL https://arxiv.org/abs/2002.10716.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pp. 5019–5031, 2018.

Sinha, A., Namkoong, H., and Duchi, J. C. Certifying some distributional robustness with principled adversarial training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Vapnik, V. N. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Wang, Y., Jha, S., and Chaudhuri, K. Analyzing the robustness of nearest neighbors to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 5120–5129, 2018.

Yang, Y., Rashtchian, C., Wang, Y., and Chaudhuri, K. Adversarial examples for non-parametric methods: Attacks, defenses and large sample limits. *CoRR*, abs/1906.03310, 2019. URL http://arxiv.org/abs/1906.03310.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R., and Chaudhuri, K. A closer look at accuracy vs. robustness, 2020.

Yin, D., Ramchandran, K., and Bartlett, P. L. Rademacher complexity for adversarially robust generalization. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7085–7094. PMLR, 2019. URL http://proceedings.mlr.press/v97/yin19b.html.