Provable Robustness of Adversarial Training for Learning Halfspaces with Noise

Difan Zou [*1] **Spencer Frei** [*2] **Quanquan Gu** [1]

Abstract

We analyze the properties of adversarial training for learning adversarially robust halfspaces in the presence of agnostic label noise. Denoting $\mathsf{OPT}_{p,r}$ as the best robust classification error achieved by a halfspace that is robust to perturbations of ℓ^p balls of radius r, we show that adversarial training on the standard binary cross-entropy loss yields adversarially robust halfspaces up to (robust) classification error $\widetilde{O}(\sqrt{\mathsf{OPT}_{2,r}})$ for $p = 2$, and $\widetilde{O}(d^{1/4}\sqrt{\mathsf{OPT}_{\infty,r}} + d^{1/2}\mathsf{OPT}_{\infty,r})$ when $p = \infty$. Our results hold for distributions satisfying anti-concentration properties enjoyed by log-concave isotropic distributions among others. We additionally show that if one instead uses a nonconvex sigmoidal loss, adversarial training yields halfspaces with an improved robust classification error of $O(\mathsf{OPT}_{2,r})$ for $p = 2$, and $O(d^{1/4}\mathsf{OPT}_{\infty,r})$ when $p = \infty$. To the best of our knowledge, this is the first work to show that adversarial training provably yields robust classifiers in the presence of noise.

1. Introduction

Modern deep learning models are powerful but brittle: standard stochastic gradient descent (SGD) training of deep neural networks can lead to remarkable performance as measured by the classification accuracy on the test set, but this performance rapidly degrades if the metric is instead *adversarially robust* accuracy. This brittleness is most apparent for image classification tasks, where neural networks trained by gradient descent achieve state-of-the-art classification accuracy on a number of benchmark tasks, but where imperceptible (adversarial) perturbations of an image can force the neural network to get nearly all of its predictions incorrect (Szegedy et al., 2014; Goodfellow et al., 2015).

[*]Equal contribution [1]Department of Computer Science, UCLA [2]Department of Statistics, UCLA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

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

To formalize the above comment, let us define the robust error of a classifier. Let $\mathcal{D}$ be a distribution over $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$, and let $f : \mathbb{R}^d \to \{\pm 1\}$ be a hypothesis classifier. For $p \in [1, \infty]$ and perturbation radius $r > 0$, the ℓ^p *robust error for radius* r is given by

$$\mathrm{err}_{\mathcal{D}}^{p,r}(f) = \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\exists \mathbf{x}' : \|\mathbf{x}-\mathbf{x}'\|_p \le r, \text{ and } y \ne f(\mathbf{x}')\big] \tag{1.1}$$

The standard accuracy of a classifier f is given by $\mathrm{err}_{\mathcal{D}}(f) = \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}(y \ne f(\mathbf{x}))$, and is equivalent to the robust accuracy at radius $r = 0$. That SGD produces neural networks f with high classification accuracy but low robust accuracy means that $\mathrm{err}_{\mathcal{D}}(f) \approx 0$ but $\mathrm{err}_{\mathcal{D}}^{p,r}(f) \approx 1$, even when r is an extremely small number.

The vulnerability of SGD-trained neural networks to adversarial examples has led researchers to introduce a number of methods aimed at improving the robustness of neural networks to adversarial examples (Kurakin et al., 2016; Madry et al., 2018; Tramèr et al., 2018; Zhang et al., 2019; Wang et al., 2019a;b). One notable approach is known as *adversarial training*, where the standard SGD algorithm is modified so that data samples are perturbed $\mathbf{x} \mapsto \mathbf{x} + \boldsymbol{\delta}$ with the aim of increasing the robust accuracy. In the same way that one minimizes the standard classification error by minimizing a surrogate loss, adversarial training seeks to minimize

$$L_{\mathcal{D}}^{p,r}(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \sup_{\mathbf{x}':\|\mathbf{x}'-\mathbf{x}\|_p\le r} \ell(yf(\mathbf{x}')), \tag{1.2}$$

where $\ell(\cdot)$ is some convex surrogate for the 0-1 loss. Unfortunately, the inner maximization problem is typically intractable, especially when f comes from a neural network function class. Indeed, it is often difficult to calculate *any* nontrivial upper bound for the robust loss $\sup_{\mathbf{x}':\|\mathbf{x}-\mathbf{x}'\|_p\le r} \ell(yf(\mathbf{x}'))$ for a fixed sample $\mathbf{x}$. A number of recent works have focused on developing upper bounds for the robust loss that are computationally tractable, which enables end-users to certify the robustness of learned classifiers by evaluating the upper bound on test samples (Raghunathan et al., 2018; Wong & Kolter, 2018; Cohen et al., 2019). Additionally, upper bounds for the robust loss can then be used as a new objective function to be minimized as an alternative to the intractable robust loss. This approach has seen impressive results in improving the adversarial robustness of classifiers, but unfortunately these

procedures do not come with a provable guarantee that the learned classifiers will be adversarially robust. To the best of our knowledge, only two works have been able to show that the standard gradient-based adversarial training of (1.2) provably yields classifiers with a guarantee on the robust (population-level) classification error: Charles et al. (2019) and Li et al. (2020). Both of these papers considered the hypothesis class of halfspaces $\mathbf{x} \mapsto \mathrm{sign}(\mathbf{w}^\top \mathbf{x})$ and assumed that the data distribution is linearly separable by a hard margin $\gamma_0 > 0$, so that for some $\mathbf{w} \in \mathbb{R}^d$, $y\mathbf{w}^\top \mathbf{x} \geq \gamma_0 > 0$ holds almost surely over $\mathcal{D}$.

In this work, we show that adversarial training provably leads to halfspaces that are approximate minimizers for the population-level robust classification error. In particular, adversarial training provably yields classifiers which are robust even when the data is not linearly separable. Let us denote the best-possible robust classification error for a halfspace as

$$\mathsf{OPT}_{p,r} = \min_{\|\mathbf{w}\|_q = 1} \mathrm{err}_{\mathcal{D}}^{p,r}(\mathbf{w}),$$

where $\mathrm{err}_{\mathcal{D}}^{p,r}(\mathbf{w})$ is the robust error induced by the halfspace classifier. Our main contributions are as follows.

1. We show that adversarial training on the robust surrogate loss (1.2) yields halfspaces with ℓ^2 robust error at most $\widetilde{O}(\sqrt{\mathsf{OPT}_{2,r}})$ when ℓ is a typical convex surrogate loss and $\mathcal{D}$ satisfies an anti-concentration property enjoyed by log-concave isotropic distributions. For $p = \infty$, our guarantee is $\widetilde{O}(d^{1/4}\sqrt{\mathsf{OPT}_{\infty,r}} + d^{1/2}\mathsf{OPT}_{\infty,r})$.

2. When ℓ is a nonconvex sigmoidal loss, the guarantees for adversarial training improves to $O\big(d^{\frac{1}{4}-\frac{1}{2p}}\|\mathbf{w}_*\|_2^{1/2}\mathsf{OPT}_{p,r}\big)$ for ℓ_p perturbations, where $\mathbf{w}^*$ of norm $\|\mathbf{w}^*\|_q = 1$ (for $1/p + 1/q = 1$) is the optimal model. This implies that adversarial training achieves $O(\mathsf{OPT}_{2,r})$ robust error for perturbations in the ℓ^2 metric, and $O(d^{1/4}\mathsf{OPT}_{\infty,r})$ robust error when $p = \infty$ in the worst case.

To the best of our knowledge, these are the first results that provide a guarantee that adversarial training will generate adversarially robust classifiers on noisy data distributions.

1.1. Additional Related Work

Adversarial training and adversarial examples have attracted significant attention recently due to the explosion of research in deep learning, but the broader problem of learning decision rules that are robust to perturbations of the data has appeared in a number of forms. One of the main motivations for support vector machines is to maximize the margin of the classifier, which can be understood as a form of robustness to perturbations of the input (Rosenblatt, 1958; Boser et al., 1992). Robust optimization is a field in its own right dedicated to the analysis of optimization algorithms that are robust to perturbations of the algorithms' inputs (Ben-Tal et al., 2009).

Following the first paper on adversarial examples in deep learning (Szegedy et al., 2014), a sequence of works sought to develop empirical methods for improving the robustness of neural network classifiers (Goodfellow et al., 2015; Papernot et al., 2016). These proposed defenses against adversarial examples were quickly defeated by more sophisticated attacks (Carlini & Wagner, 2017). This led a number of authors to develop *certifiable* defenses against adversarial attacks, where one can prove that the defense algorithm will be robust to adversarial perturbations (Wong & Kolter, 2018; Raghunathan et al., 2018). These works typically derive an upper bound for the robust loss that can be computed exactly and then introduce optimization procedures for minimizing the upper bound. This allows for one to certify whether or not a classifier is provably robust to adversarial perturbations for a given sample. But since the procedure is based upon minimizing an upper bound for the desired error, there is no guarantee that every classifier which is trained using this procedure will (provably) yield a classifier that has nontrivial robust classification accuracy.

In terms of provable guarantees for learning adversarially robust classifiers, adversarial training was shown to yield provably robust halfspace classifiers by Charles et al. (2019) and Li et al. (2020) under the assumption that there exists a robust classifier with perfect accuracy that separates the data by a large margin. A separate approach for developing robust classifiers is known as randomized smoothing (Salman et al., 2019; Lécuyer et al., 2019; Cohen et al., 2019), where one can convert a base classifier into a robust classifier by smoothing out the predictions of the base classifier over Gaussian noise perturbations of the input. Nandy et al. (2021) showed that test-time adaptive batch normalization can lead to certifiably robust classifiers. Gao et al. (2019) and Zhang et al. (2020) showed that adversarial training with multilayer neural networks leads to classifiers with small robust training loss, but were not able to translate these into guarantees for small test (population-level) robust error. Montasser et al. (2020) showed that the standard gradient descent algorithm on the (non-robust) empirical risk using a convex margin loss yields halfspaces that are robust in the presence of random classification noise.[1] Diakonikolas et al. (2020a) studied the computational complexity of learning

[1] Random classification noise (RCN) is a generalization of the realizable setting, where an underlying halfspace $y = \mathrm{sign}(\mathbf{w}^\top \mathbf{x})$ has labels flipped with probability p. By contrast, in the adversarial label noise setting we consider in this paper, one makes no assumptions on the relationship between $\mathbf{x}$ and y.

robust halfspaces in the agnostic noise setting.

We wish to emphasize that in this work we are interested in developing *computationally efficient* algorithms for learning adversarially robust halfspaces in the presence of noise. (Cullina et al., 2018) recently developed a notion of adversarial VC dimension, which allows for a characterization of the number of samples necessary to learn robust classifiers in the presence of noise by analyzing the robust empirical risk minimizer (ERM). However, the non-convexity of the zero-one loss makes the task of finding a robust ERM a highly non-trivial task. Indeed, it is known that no polynomial time algorithm can agnostically learn standard (non-robust) halfspaces up to risk $O(\mathsf{OPT}_{p,0}) + \varepsilon$ without distributional assumptions (Daniely, 2016), although standard VC dimension arguments show that $\mathrm{poly}(d, \varepsilon^{-1})$ samples suffice for the ERM to achieve $\mathsf{OPT}_{p,0} + \varepsilon$ risk. Thus, in order to develop computationally efficient algorithms that can robustly learn up to robust risk $O(\mathsf{OPT}_{p,r})$, we must make assumptions on the distribution.

There are a number of other important questions in adversarial robustness for which a detailed review is beyond the scope of this paper. We briefly note that some related topics include understanding the possible tradeoffs between robust accuracy and non-robust accuracy (Zhang et al., 2019; Tsipras et al., 2019; Javanmard et al., 2020; Raghunathan et al., 2020; Yang et al., 2020; Wu et al., 2020); what types of features robust classifiers depend upon (Ilyas et al., 2019); and the transferability of robust classifiers (Salman et al., 2020).

1.2. Notation

We use bold-faced letters to denote vectors. For a scalar x, we use $\mathrm{sgn}(x) \in \{+1, -1\}$ to denote its sign. For $p \in [1, \infty]$, we denote $\mathcal{B}_p(\mathbf{x}, r) = \{\mathbf{x}' : \|\mathbf{x} - \mathbf{x}'\|_p \le r\}$ as the ℓ^p ball of radius r centered at $\mathbf{x}$. We use $\mathcal{S}_q^{d-1}$ to denote the unit ℓ_q sphere. Given two vectors $\mathbf{w}$ and $\mathbf{v}$, we use $\angle(\mathbf{w}, \mathbf{v})$ to denote the angle between these two vectors. We use the indicator function $\mathbb{1}(\mathfrak{E})$ to denote 1 on the event $\mathfrak{E}$ and 0 elsewhere. We use the standard $O(\cdot)$ and $\Omega(\cdot)$ notations to hide universal constants, with $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ additionally ignoring logarithmic factors. The notation $g(x) = \Theta(f(x))$ denotes a function with growth rate satisfying both $g(x) = O(f(x))$ and $g(x) = \Omega(f(x))$.

1.3. Paper Organization

The remainder of the paper is organized as follows. In Section 2, we describe our guarantees for adversarial training on convex loss functions. In Section 3, we show that by using a nonconvex sigmoidal loss, we can achieve improved guarantees for the robust classification accuracy of halfspaces. We conclude in Section 4.

Algorithm 1 Adversarial Training

input: Training dataset $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1,\dots,n}$, step size η
for $k = 0, 1, \dots, K$ **do**
  for $i = 1, \dots, n$ **do**
    $\boldsymbol{\delta}_i^{(k)} := \mathrm{argmax}_{\|\boldsymbol{\delta}\|_p \le r} \ell(y_i \mathbf{w}_k^\top (\mathbf{x}_i + \boldsymbol{\delta}))$
  end for
  $\mathbf{w}_{k+1} = \mathbf{w}_k - \frac{\eta}{n} \sum_{i=1}^n \ell'(y_i w^\top (\mathbf{x}_i + \boldsymbol{\delta}_i^{(k)})) y_i (\mathbf{x}_i + \boldsymbol{\delta}_i^{(k)})$
end for
output: $\{\mathbf{w}_k\}_{k=0,\dots,K}$

2. Adversarial Training with Convex Surrogates

Our first set of results is for the case that the loss function ℓ appearing in the definition of the robust loss (1.2) is a typical decreasing convex surrogates of the zero-one loss, such as the cross entropy $\ell(z) = \log(1 + \exp(-z))$ or hinge loss $\ell(z) = \max(0, 1 - z)$. We consider a standard approach for gradient descent-based adversarial training of the objective (1.2), which consists of two parts: (1) an inner maximization, and (2) an outer minimization. For the inner maximization, we find the optimal perturbation of the input which maximizes $\ell(y\mathbf{w}^\top(\mathbf{x} + \boldsymbol{\delta}))$ for $\boldsymbol{\delta} \in \mathcal{B}_p(0, r)$. For more complicated model classes, such as neural networks, the inner maximization procedure can often be very difficult to optimize. As such, it is usually difficult to derive provable guarantees for the robustness of adversarial training procedures. However, in the linear model class that we consider, we can solve the inner maximization procedure exactly. When ℓ is decreasing, this maximization problem is equivalent to

$$\arg \min_{\|\boldsymbol{\delta}\|_p \le r} y\mathbf{w}^\top(\mathbf{x} + \boldsymbol{\delta}).$$

Using calculus we can solve for the exact solution to this minimization problem. The optimal perturbation is given by $\boldsymbol{\delta}^* = \boldsymbol{\delta}^*(\mathbf{w}, r, y)$, with components

$$\delta_j^* = -ry \cdot \mathrm{sgn}(w_j)|w_j|^{q-1} / \|\mathbf{w}\|_q^{q-1}, \qquad (2.1)$$

where q is the Hölder conjugate to p so that $1/q + 1/p = 1$. The ability to solve the inner maximization procedure exactly means that the only remaining part is to solve the outer minimization. For this, we use the standard gradient descent algorithm on the perturbed examples. We note that we do not differentiate through the samples in the gradient updates—although the perturbed examples depend on the weights (via $\boldsymbol{\delta}^*$), we treat these perturbed samples as if they are independent of $\mathbf{w}$. The update rule is explicitly given in Algorithm 1.

Our first result is that Algorithm 1 efficiently minimizes the robust empirical risk.

Lemma 2.1. Assume ℓ is convex, decreasing, and 1-Lipschitz. Let $\mathbf{w}^* \in \mathbb{R}^d$ be arbitrary. Let $p \in [1, \infty]$, and assume that $\|\mathbf{x}\|_p \leq 1$ a.s. If $p \leq 2$, let $H = 4$, and if $p > 2$, let $H = 4d$. Let $\varepsilon > 0$ and be arbitrary. If $\eta \leq \varepsilon H^{-1}/4$, then for any initialization $\mathbf{w}_0$, if we denote $\mathbf{w}_k$ as the k-th iterate of Algorithm 1, by taking $K = \varepsilon^{-1}\eta^{-1}\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2$, we have there exists a $k^* \leq K$ such that $\|\mathbf{w}_{k^*} - \mathbf{w}^*\|_2^2 \leq \|\mathbf{w}_0 - \mathbf{w}^*\|_2^2$ and

$$L_{\mathcal{S}}^{p,r}(\mathbf{w}_{k^*}) \leq L_{\mathcal{S}}^{p,r}(\mathbf{w}^*) + \varepsilon.$$

The proof for the above Lemma can be found in Appendix B.1. To convert the guarantee for the empirical risk into one for the population risk, we will utilize an argument based on robust Rademacher complexity (Yin et al., 2019). This is possible because Lemma 2.1 shows that the weights returned by Algorithm 1 stay in a norm-bounded region.

Lemma 2.2 (Population robust loss)**.** Assume that $\|\mathbf{x}\|_p \leq 1$ a.s. and that ℓ is convex, decreasing, and 1-Lipschitz. Let $\mathbf{w}^* \in \mathbb{R}^d$ be such that $\|\mathbf{w}^*\|_q \leq \rho$ for some $\rho > 0$, and denote $B = \ell(0) + 2d^{|1/q-1/2|}(1+r)\rho$ and $\bar{B} = 2d^{|1/q-1/2|}B$. Denote $\mathfrak{R}_p = n^{-1}\mathbb{E}_\sigma\left[\|\sum_{i=1}^n \sigma_i \mathbf{x}_i\|_p\right]$. Then for any $\varepsilon > 0$, using the same notation from Lemma 2.1, running Algorithm 1 with $\mathbf{w}_0 = 0$ ensures that there exists $k^* \leq K = \max\{1, d^{1/q-1/2}\}\eta^{-1}\varepsilon^{-1}\rho^2$ such that with probability at least $1-\delta$,

$$L_{\mathcal{D}}^{p,r}(\mathbf{w}_{k^*}) \leq L_{\mathcal{D}}^{p,r}(\mathbf{w}^*) + \epsilon + 4\bar{B}\rho\mathfrak{R}_p + 4\bar{B}\frac{\rho r}{\sqrt{n}} + 6B\sqrt{\frac{\log(2K/\delta)}{2n}}$$

The proof for Lemma 2.2 is in Appendix B.2. We note that the term $\mathfrak{R}_p$ is a common complexity term that takes the form $O(1/\sqrt{n})$ for $p = 2$ and $O(\log(d)/\sqrt{n})$ for $p = \infty$; see e.g. Lemmas 26.10 and 26.11 of Shalev-Shwartz & Ben-David (2014).

Now that we have shown that adversarial training yields hypotheses which minimize the *surrogate* robust risk $L_{\mathcal{D}}^{p,r}$, the next step is to show that this minimizes the robust classification error $\mathrm{err}_{\mathcal{D}}^{p,r}$. (Since $L_{\mathcal{D}}^{p,r}$ is only an upper bound for $\mathrm{err}_{\mathcal{D}}^{p,r}$, minimizers for $L_{\mathcal{D}}^{p,r}$ do not necessarily minimize $\mathrm{err}_{\mathcal{D}}^{p,r}$.) Recently, Frei et al. (2020) introduced the notion of soft margins in order to translate minimizers of surrogate losses to approximate minimizers for classification error, and we will use a similar approach here. Let us first define soft margin functions.

Definition 2.3. Let $q \in [1, \infty]$. Let $\bar{\mathbf{v}} \in \mathbb{R}^d$ satisfy $\|\bar{\mathbf{v}}\|_q = 1$. We say $\bar{\mathbf{v}}$ satisfies the ℓ^q *soft margin condition with respect to a function* $\phi_{\bar{\mathbf{v}},q} : \mathbb{R} \to \mathbb{R}$ if for all $\gamma \in [0, 1]$, it holds that

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_x}\left[\mathbb{1}\left(|\bar{\mathbf{v}}^\top \mathbf{x}| \leq \gamma\right)\right] \leq \phi_{\bar{\mathbf{v}},q}(\gamma).$$

The properties of the ℓ^q soft margin function for $q = 2$ for a variety of distributions were shown by Frei et al. (2020). We collect some of these in the examples below, but let us first introduce the following definitions which will be helpful for understanding the soft margin.

Definition 2.4. For $\bar{\mathbf{v}}, \bar{\mathbf{v}}' \in \mathbb{R}^d$, denote by $p_{\bar{\mathbf{v}},\bar{\mathbf{v}}'}(\cdot)$ the marginal distribution of $\mathbf{x} \sim \mathcal{D}_x$ on the subspace spanned by $\bar{\mathbf{v}}$ and $\bar{\mathbf{v}}'$. We say $\mathcal{D}_x$ satisfies U*-anti-concentration* if there is some $U > 0$ such that for any two vectors $\bar{\mathbf{v}}, \bar{\mathbf{v}}'$ satisfying $\|\bar{\mathbf{v}}\|_2 = \|\bar{\mathbf{v}}'\|_2 = 1$, we have $p_{\bar{\mathbf{v}},\bar{\mathbf{v}}'}(\mathbf{z}) \leq U$ for all $\mathbf{z} \in \mathbb{R}^2$. We say that $\mathcal{D}_x$ satisfies (U', R)*-anti-anti-concentration* if there exists $U', R > 0$ such that $p_{\bar{\mathbf{v}},\bar{\mathbf{v}}'}(\mathbf{z}) \geq 1/U'$ for all $\mathbf{z} \in \mathbb{R}^2$ satisfying $\|\mathbf{z}\|_2 \leq R$.

Anti-concentration and anti-anti-concentration have recently been used for deriving guarantees agnostic PAC learning guarantees for learning halfspaces (Diakonikolas et al., 2020b;c; Frei et al., 2020). Log-concave isotropic distributions, such as the standard Gaussian in d dimensions or the uniform distribution over any convex set, satisfy U-anti-concentration and (U', R)-anti-anti-concentration with each of U, U', and R being universal constants independent of the dimension of the input space. Below, we collect some of the properties of the ℓ^2 soft margin function.

Example 2.5.

1. For any $q \in [1, \infty]$, if $\bar{\mathbf{v}} \in \mathbb{R}^d$ satisfies $\|\bar{\mathbf{v}}\|_q = 1$ and $|\bar{\mathbf{v}}^\top \mathbf{x}| > \gamma^*$ a.s., then $\phi_{\bar{\mathbf{v}},q}(\gamma) = 0$ for $\gamma < \gamma^*$.
2. If $\mathcal{D}_x$ satisfies U-anti-concentration, then $\phi_{\bar{\mathbf{v}},2}(\gamma) = O(\gamma)$.
3. If $\mathcal{D}_x$ satisfies (U', R)-anti-anti-concentration, then for $\gamma \leq R$, $\phi_{\bar{\mathbf{v}},2}(\gamma) = \Omega(\gamma)$ holds.
4. Isotropic log-concave distributions (i.e. isotropic distributions with log-concave probability density functions) satisfy U-anti-concentration and (U', R)-anti-anti-concentration for $U, U', R = \Theta(1)$.

Proofs for these properties can be found in Appendix B.3. For $q \neq 2$, the soft margin function will depend upon the ratio of the ℓ^q to the ℓ^2 norm, since we have the identity, for any $\bar{\mathbf{v}}$ satisfying $\|\bar{\mathbf{v}}\|_q = 1$,

$$\begin{aligned}\phi_{\bar{\mathbf{v}},q}(\gamma) &= \mathbb{P}(|\bar{\mathbf{v}}^\top \mathbf{x}| \leq \gamma) \\ &= \mathbb{P}\left(\frac{\bar{\mathbf{v}}^\top}{\|\bar{\mathbf{v}}\|_2} \leq \frac{\gamma}{\|\bar{\mathbf{v}}\|_2}\right) \\ &= \phi_{\bar{\mathbf{v}}/\|\bar{\mathbf{v}}\|_2,2}(\gamma/\|\bar{\mathbf{v}}\|_2). \end{aligned} \tag{2.2}$$

Thus, the ℓ^q soft margin function scales with the ratio of $\|\bar{\mathbf{v}}\|_q/\|\bar{\mathbf{v}}\|_2$. The case $q = 1$ corresponds to the ℓ^∞ perturbation and is of particular interest. By Cauchy–Schwarz, $\|\bar{\mathbf{v}}\|_1 \leq \sqrt{d}\|\bar{\mathbf{v}}\|_2$, and this bound is tight in the worst case (take $\bar{\mathbf{v}} = \mathbf{1} \in \mathbb{R}^d$). Thus the ℓ^1 soft margin has an unavoidable dimension dependence in the worst case. We collect

the above observations, with some additional properties that we show in Appendix B.3, in the following example.

Example 2.6. 1. If $\mathcal{D}_x$ satisfies U-anti-concentration, and if $q \in [1,2]$, then for any $\bar{\mathbf{v}} \in \mathbb{R}^d$ with $\|\bar{\mathbf{v}}\|_q = 1$, $\phi_{\bar{\mathbf{v}},q}(\gamma) = O(\gamma d^{\frac{1}{q}-\frac{1}{2}})$.

2. If $\mathcal{D}_x$ satisfies (U', R)-anti-anti-concentration and if $q \in [1,2]$, then for any $\bar{\mathbf{v}} \in \mathbb{R}^d$ with $\|\bar{\mathbf{v}}\|_q = 1$, for $\gamma \le \Theta(R)$, it holds that $\phi_{\bar{\mathbf{v}},q}(\gamma) = \Omega(\gamma)$.

We now can proceed with relating the minimizer of the surrogate loss $L_{\mathcal{D}}^{p,r}$ to that of $\mathrm{err}_{\mathcal{D}}^{p,r}$ by utilizing the soft margin. The proof for the following Lemma is in Appendix B.4.

Lemma 2.7. Let $p, q \in [1,\infty]$ be such that $1/p + 1/q = 1$ and assume $\|\mathbf{x}\|_p \le 1$ a.s. Let $\bar{\mathbf{v}} := \min_{\|\mathbf{w}\|_q=1} \mathrm{err}_{\mathcal{D}}^{p,r}(\mathbf{w})$, so that $\mathrm{err}_{\mathcal{D}}^{p,r}(\bar{\mathbf{v}}) = \mathsf{OPT}$. Assume that $\|\mathbf{x}\|_p \le 1$ a.s. For $\rho > 0$, denote $\mathbf{v} := \rho\bar{\mathbf{v}}$ as a scaled version of the population risk minimizer for $\mathrm{err}_{\mathcal{D}}^{p,r}(\cdot)$. Assume ℓ is 1-Lipschitz, non-negative and decreasing. Then we have

$$L_{\mathcal{D}}^{p,r}(\mathbf{v}) \le \inf_{\gamma>0} \Big\{ (\ell(0)+\rho)\mathsf{OPT}_{p,r} + \ell(0)\phi_{\bar{\mathbf{v}},q}(r+\gamma) + \ell(\rho\gamma) \Big\}. \tag{2.3}$$

Thus, if $\ell(0) > 0$,

$$\mathrm{err}_{\mathcal{D}}^{p,r}(\mathbf{v}) \le [\ell(0)]^{-1} \inf_{\gamma>0} \Big\{ (\ell(0)+\rho)\mathsf{OPT}_{p,r} + \ell(0)\phi_{\bar{\mathbf{v}},q}(r+\gamma) + \ell(\rho\gamma) \Big\}.$$

Using Lemmas 2.7 and 2.2, we can derive the following guarantee for the robust classification error for classifiers learned using adversarial training.

Theorem 2.8. Suppose $\ell \ge 0$ is convex, decreasing, and 1-Lipschitz. Let $p \in [1,\infty]$ and $q \in [1,\infty]$ satisfy $1/p + 1/q = 1$. Denote $H = 4$ if $p \le 2$ and $H = 4d$ if $p > 2$. Let $\varepsilon > 0$ be arbitrary, and fix $\eta \le \varepsilon H^{-1}/4$. For any $\gamma > 0$, running Algorithm 1 with $\mathbf{w}_0 = 0$ for $K = \max\{1, d^{\frac{2}{q}-1}\}\varepsilon^{-1}\eta^{-1}\ell^{-2}(1/\varepsilon)\gamma^{-2}$ iterations, with probability at least $1-\delta$, there exists $k^* \le K$ such that

$$\begin{aligned}\mathrm{err}_{\mathcal{D}}^{p,r}(\mathbf{w}_{k^*}) &\le \big(1 + [\ell(0)]^{-1}\cdot\ell^{-1}(1/\varepsilon)\cdot\gamma^{-1}\big)\,\mathsf{OPT}_{p,r} \\ &+ \phi_{\bar{\mathbf{v}},q}(r+\gamma) + [\ell(0)]^{-1}\varepsilon + 4[\ell(0)]^{-1}\bar{B}\gamma^{-1}\ell^{-1}(1/\varepsilon)\mathfrak{R}_p \\ &+ [\ell(0)]^{-1}\left[\frac{4\bar{B}\gamma^{-1}\ell^{-1}(1/\varepsilon)r}{\sqrt{n}} + 6B\sqrt{\frac{\log(2K/\delta)}{n}}\right],\end{aligned}$$

where $B = \ell(0) + 2d^{|2-q|/2}(1+r)\gamma^{-1}\ell^{-1}(\varepsilon)$, $\mathfrak{R}_p = n^{-1}\mathbb{E}_{\sigma_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\pm 1)}[\|\sum_{i=1}^n \sigma_i \mathbf{x}_i\|_p]$, and $\bar{B} = 2d^{|2-q|/2}B$.

Proof. The result follows by using Lemmas 2.2 and 2.7 with the choice of $\rho = \gamma^{-1}\ell^{-1}(1/\varepsilon)$. □

In order to realize the right-hand-side of the above bound for the robust classification error, we will need to analyze the properties of the soft margin function $\phi_{\bar{\mathbf{v}},q}$ and then optimize over γ. We will do so in the following corollaries. We start by considering hard margin distributions.

Corollary 2.9 (Hard margin)**.** Let $p \ge 2$ and $q \in [1,2]$ be such that $1/p + 1/q = 1$. Assume $\|\mathbf{x}\|_p \le 1$ a.s. Suppose $\mathbf{v}^*$ is such that $\|\mathbf{v}^*\|_q = 1$ and for some $\gamma_0 \in [0,1]$, $|\langle \mathbf{v}^*, \mathbf{x}\rangle| \ge \gamma_0$, and $\mathrm{err}_{\mathcal{D}}^{p,r}(\mathbf{v}^*) = \min_{\|\mathbf{w}\|_q=1} \mathrm{err}_{\mathcal{D}}^{p,r}(\mathbf{w}) = \mathsf{OPT}_{p,r}$. Suppose we consider the perturbation radius $r = (1-\nu)\gamma_0$ for some $\nu \in (0,1)$. Consider the cross entropy loss for simplicity, and let $\eta \le \mathsf{OPT}_{p,r}H^{-1}/4$, where $H = 4$ if $p \le 2$ and $H = 4d$ if $p > 2$. Then the adversarial training in Algorithm 1 started from $\mathbf{w}_0 = 0$ finds classifiers satisfying $\mathrm{err}_{\mathcal{D}}^{p,r}(\mathbf{w}_k) = \widetilde{O}(\nu^{-1}\gamma_0^{-1}\mathsf{OPT}_{p,r})$ within $K = \widetilde{O}(\eta^{-1}d^{\frac{2}{p}-1}\gamma_0^{-2}\nu^{-2}\mathsf{OPT}_{p,r}^{-1})$ iterations provided $n = \widetilde{\Omega}(\gamma_0^{-4}\nu^{-2}\mathsf{OPT}_{p,r}^{-2})$.

Proof. We sketch the proof here and leave the detailed calculations for Appendix B.5. By the definition of soft margin, $\phi_{\bar{\mathbf{v}}^*,q}(\gamma_0) = 0$, and so if we choose $\gamma = \nu\gamma_0$ and $\varepsilon = \mathsf{OPT}_{p,r}$ in Theorem 2.8, we get a bound for the robust classification error of the form $\widetilde{O}(\nu^{-1}\gamma_0^{-1}\mathsf{OPT}_{p,r}) + \widetilde{O}(1)\cdot\mathfrak{R}_p + \widetilde{O}(1/\sqrt{n})$ by using the fact that $\ell^{-1}(1/\varepsilon) = O(\log(1/\varepsilon))$ for the cross entropy loss. Standard arguments in Rademacher complexity show that $\mathfrak{R}_p = \widetilde{O}(1/\sqrt{n})$, completing the proof. □

The above corollary shows that if the best classifier separates the samples with a hard margin of γ_0 (including when it makes incorrect predictions), then adversarial training will produce a classifier that has robust classification error within a constant factor of the best-possible robust classification error. This can be seen as a generalization of the results of Charles et al. (2019) and Li et al. (2020) from distributions that can achieve perfect robust classification accuracy (with a hard margin) to ones where significant label noise can be present.

Our next result is for the class of distributions satisfying the anti-concentration properties described in Definition 2.4.

Corollary 2.10 (Anti-concentration distributions)**.** Let $p \in [2,\infty]$ and assume $\|\mathbf{x}\|_p \le 1$ a.s. Suppose $\mathcal{D}_x$ satisfies U-anti-concentration and (U', R)-anti-anti-concentration for $U, U', R = \Theta(1)$. Consider the cross entropy loss for simplicity, and let $\eta \le \mathsf{OPT}_{p,r}H^{-1}/4$, where $H = 4$ if $p \le 2$ and $H = 4d$ if $p > 2$. Then for perturbations satisfying $r \le R$, the adversarial training in Algorithm 1 started from $\mathbf{w}_0 = 0$ finds classifiers satisfying

$$\mathrm{err}_{\mathcal{D}}^{p,r}(\mathbf{w}_k) = \widetilde{O}\big(d^{\frac{1}{4}-\frac{1}{2p}}\sqrt{\mathsf{OPT}_{p,r}} + d^{\frac{1}{2}-\frac{1}{p}}\mathsf{OPT}_{p,r}\big),$$

within $K = \widetilde{O}(\eta^{-1}d^{\frac{3}{2p}-\frac{3}{4}}\mathsf{OPT}_{p,r}^{-3})$ iterations provided $n = \widetilde{\Omega}(d^{1-2p}\mathsf{OPT}_{p,r}^{-2})$.

Proof. We again sketch the proof here and leave the detailed calculations for Appendix B.5. Example 2.6 shows that $\phi_{\bar{\mathbf{v}}^*,q}(a) = O(ad^{\frac{1}{q}-\frac{1}{2}}) = O(ad^{\frac{1}{2}-\frac{1}{p}})$. Anti-anti-concentration can be shown to imply that $r = O(\mathsf{OPT}_{p,r})$, and thus $\phi_{\bar{\mathbf{v}}^*,q}(\gamma + r) = O(\gamma d^{\frac{1}{2}-\frac{1}{p}}) + O(d^{\frac{1}{2}-\frac{1}{p}}\mathsf{OPT}_{p,r})$. The first term is of the same order as $\gamma^{-1}\mathsf{OPT}_{p,r}$ when $\gamma = \mathsf{OPT}_{p,r}^{1/2} d^{\frac{1}{2p}-\frac{1}{4}}$, and results in a term of the form $\widetilde{O}(d^{\frac{1}{4}-\frac{1}{2p}}\sqrt{\mathsf{OPT}_{p,r}})$. The other terms following using an argument similar to that of Corollary 2.9. □

The above shows that adversarial training yields approximate minimizers for the robust classification accuracy for halfspaces over distributions satisfying anti-concentration assumptions. In particular, this result holds for any log-concave isotropic distribution, such as the standard Gaussian or the uniform distribution over a convex set.

Remark 2.11. We note that although the guarantees in this section are for (full-batch) gradient descent-based adversarial training, nearly identical guarantees can be also derived for online SGD-based adversarial training. We give the details on this extension in Appendix C.

3. Adversarial Training with Nonconvex Sigmoidal Loss

We now show that if instead of using a typical convex loss function we use a particular nonconvex sigmoidal loss, we can improve our guarantees for the robust classification error when using adversarial training. We note that the approach of using nonconvex loss functions to derive improved guarantees for learning halfspaces with agnostic label noise was first used by Diakonikolas et al. (2020c). Our results in this section will rely upon the following assumption on the distribution $\mathcal{D}_x$.

Assumption 3.1. 1. $\mathcal{D}_x$ is mean zero and isotropic, i.e. its covariance matrix is the identity.

2. $\mathcal{D}_x$ satisfies U-anti-concentration and (U', R)-anti-anti-concentration, where $U, U', R = \Theta(1)$.

The loss function we consider is defined by

$$\ell(z) = e^{-z/\sigma} \cdot \mathbb{1}(z > 0) + (2 - e^{z/\sigma}) \cdot \mathbb{1}(z \le 0), \quad (3.1)$$

where $\sigma > 0$ is a scalar factor to be specified later. In addition to using the loss function (3.1), we additionally scale the weight vector, so that the surrogate loss we consider in this section is is

$$L_{\mathcal{D}}^{p,r}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\sup_{\mathbf{x}'\in\mathcal{B}_p(\mathbf{x},r)} \ell\left(\frac{y\mathbf{w}^\top\mathbf{x}'}{\|\mathbf{w}\|_q}\right)\right]$$

The adversarial training algorithm that we consider for the loss function (3.1) is a variant of Algorithm 1, where we introduce a projection step to normalize the weights after each gradient update. We additionally use the online stochastic gradient descent algorithm as opposed to full-batch gradient descent. For this reason we call the algorithm we use for learning halfspaces that are robust to ℓ^p perturbations of radius r by the name $\mathsf{PSAT}(p, r)$, which we describe in Algorithm 2. We note that when $p = \infty$ or $p = 2$ (i.e., $q = 1$ or $q = 2$, resp.), the projection can be done efficiently in $O(d)$ (Duchi et al., 2008) or $O(1)$ time respectively.

Algorithm 2 Projected Stochastic Adversarial Training ($\mathsf{PSAT}(p, r)$)

input: initial model parameter $\mathbf{w}_1$ with $\|\mathbf{w}_1\|_q = 1$, learning rate η, perturbation limit r.
for $k = 1, \ldots, K$ **do**
  Query data $(\mathbf{x}_k, y_k)$ from data distribution $\mathcal{D}$
  $\boldsymbol{\delta}_k := \operatorname{argmax}_{\|\boldsymbol{\delta}\|_p \le r} \ell\left(\frac{y_k \mathbf{w}_k^\top(\mathbf{x}_k+\boldsymbol{\delta})}{\|\mathbf{w}_k\|_q}\right)$
  Update $\widehat{\mathbf{w}}_{k+1} \leftarrow \mathbf{w}_k - \eta\nabla\ell\left(\frac{y_k \mathbf{w}_k^\top(\mathbf{x}_k+\boldsymbol{\delta}_k)}{\|\mathbf{w}_k\|_q}\right)$
  Project $\mathbf{w}_{k+1} \leftarrow \arg\min_{\mathbf{w}:\|\mathbf{w}\|_q=1} \|\widehat{\mathbf{w}}_{k+1} - \mathbf{w}\|_2$
end for
output: $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K$

In the below theorem we describe our guarantees for the robust classification error of halfspaces learned using Algorithm 2.

Theorem 3.2. Suppose the data distribution $\mathcal{D}$ satisfies Assumption 3.1. Let $\sigma = r$ and $\mathbf{w}^* = \arg\min_{\|\mathbf{w}\|_q=1} \mathrm{err}_{\mathcal{D}}^{p,r}(\mathbf{w})$ be the optimal model such that $\mathrm{err}_{\mathcal{D}}^{p,r}(\mathbf{w}^*) = \mathsf{OPT}_{p,r}$. If $\mathrm{err}_{\mathcal{D}}(\mathbf{w}^*) = O(rd^{2/p-1})$ and $r = O\big(d^{\frac{3}{2p}-\frac{3}{4}}\big)$, then running the adversarial training algorithm $\mathsf{PSAT}(p, r)$ for $K = O\big(d\|\mathbf{w}_1 - \mathbf{w}^*\|_2^2 \delta^{-2} r^{-4} d^{\frac{1}{2}-\frac{1}{p}}\big)$ iterations, with probability at least $1 - \delta$, there exists a $k^* \le K$ such that

$$\mathrm{err}_{\mathcal{D}}^{p,r}(\mathbf{w}_{k^*}) = O\big(d^{\frac{1}{4}-\frac{1}{2p}} \cdot \|\mathbf{w}^*\|_2^{1/2} \cdot \mathsf{OPT}_{p,r}\big).$$

We note that the robust classification error achieved by adversarial training depends on the ℓ_2 norm of the optimizer $\mathbf{w}^*$, which satisfies $d^{1/2-1/p} \le \|\mathbf{w}^*\|_2 \le 1$ since $\|\mathbf{w}^*\|_q = 1$ (where $1/p + 1/q = 1$). The strongest guarantees arise when $\|\mathbf{w}^*\|_2 = d^{1/2-1/p}$, which results in a robust classification error guarantee of $O(\mathsf{OPT}_{p,r})$, while in the worst case $\|\mathbf{w}^*\|_2 = 1$ and the guarantee is $O(d^{\frac{1}{4}-\frac{1}{2p}}\mathsf{OPT}_{p,r})$ robust error. Note that for ℓ^2 perturbations, $\|\mathbf{w}^*\|_2 = 1$ and so our guarantee is always $O(\mathsf{OPT}_{2,r})$.

In the remainder of this section we will prove Theorem 3.2. A key quantity in our proof is the inner product $\mathbf{w}^{*\top}\nabla L_{\mathcal{D}}^{p,r}(\mathbf{w})$, where $\mathbf{w}^*$ is the optimal robust halfspace classifier.[2] To get an idea for why this quantity is important, consider the gradient flow approach to minimizing

[2]Here we slightly abuse the notation since in fact the gradient

$\|\mathbf{w}(t)-\mathbf{w}^*\|_2^2$,

$$\frac{\mathrm{d}\|\mathbf{w}(t)-\mathbf{w}^*\|_2^2}{\mathrm{d}t} = -\langle \mathbf{w}(t)-\mathbf{w}^*, \nabla L_{\mathcal{D}}^{p,r}(\mathbf{w}(t))\rangle \quad (3.2)$$

If we denote $h(\mathbf{w},\mathbf{x}) = \mathbf{w}^\top\mathbf{x}/\|\mathbf{w}\|_q$, then we have the identity

$$\nabla L_{\mathcal{D}}^{p,r}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)}\left[\ell'\left(yh(\mathbf{w},\mathbf{x}+\boldsymbol{\delta})\right) y\nabla_{\mathbf{w}} h(\mathbf{w},\mathbf{x}+\boldsymbol{\delta})\right],$$

where

$$\nabla_{\mathbf{w}} h(\mathbf{w},\mathbf{x}+\boldsymbol{\delta}) = \left(\mathbf{I} - \frac{\bar{\mathbf{w}}\mathbf{w}^\top}{\|\mathbf{w}\|_q^q}\right)\frac{\mathbf{x}+\boldsymbol{\delta}}{\|\mathbf{w}\|_q}, \quad (3.3)$$

where we denote the vector $\bar{\mathbf{w}}$ as having components $\bar{w}_j = |w_j|^{q-1}\mathrm{sgn}(w_j)$. Then we have $\mathbf{w}^\top\nabla_{\mathbf{w}} h(\mathbf{w},\mathbf{x}+\boldsymbol{\delta}) = 0$ since $\mathbf{w}^\top\bar{\mathbf{w}} = \|\mathbf{w}\|_q^q$. In particular, substituting this into (3.2), we get

$$\frac{\mathrm{d}\|\mathbf{w}(t)-\mathbf{w}^*\|_2^2}{\mathrm{d}t} = \mathbf{w}^{*\top}\nabla L_{\mathcal{D}}^{p,r}(\mathbf{w}(t)). \quad (3.4)$$

This implies that the more negative the quantity $\mathbf{w}^{*\top}\nabla L_{\mathcal{D}}^{p,r}(\mathbf{w}(t))$ is, the faster the iterates of gradient flow will converge to $\mathbf{w}^*$.

The key, then, is to derive bounds on the quantity

$$\mathbf{w}^{*\top}\nabla L_{\mathcal{D}}^{p,r}(\mathbf{w}) = \mathbf{w}^{*\top}\mathbb{E}_{(\mathbf{x},y)}\left[\ell'\left(yh(\mathbf{w},\mathbf{x}+\boldsymbol{\delta})\right) y\nabla_{\mathbf{w}} h(\mathbf{w},\mathbf{x}+\boldsymbol{\delta})\right],$$

where $\mathbf{w}$ is an arbitrary vector which we will take to be the iterates of Algorithm 2. The challenge here is that the prescence of agnostic label noise means there are no *a priori* relationships between $\mathbf{x}$ and y, making it unclear how to deal with the appearance of both of these terms in the expectation. To get around this, we will use a similar high-level idea as did Diakonikolas et al. (2020c), in which we swap the label y with the prediction of the optimal solution $\mathbf{w}^*$. Then the inner product $\mathbf{w}^{*\top}\nabla L_{\mathcal{D}}^{p,r}(\mathbf{w})$ can be upper bounded only using the information of $\mathbf{w}$, $\mathbf{w}^*$, the distribution of $\mathbf{x}$, and the classification error $\mathrm{err}_{\mathcal{D}}(\mathbf{w}^*)$. The details of this calculation become more complicated since adversarial training introduces perturbations that also depend on the label: the optimal perturbation for $\mathbf{w}$ is $-ry\bar{\mathbf{w}}/\|\mathbf{w}\|_q^{q-1}$. This requires additional attention in the proof. Finally, since we consider general ℓ^p perturbations, the normalization by norms with $p \neq 2$ introduces additional complications.

Let us begin with some basic calculations. We first give some general calculations which will be frequently used in the subsequent analyses. Let $h(\mathbf{w},\mathbf{x}) = \mathbf{w}^\top\mathbf{x}/\|\mathbf{w}\|_q$ be the prediction of the normalized classifier and denote the event

$$S = \{(\mathbf{x},y) : y = \mathrm{sgn}(\mathbf{w}^{*\top}\mathbf{x})\}, \quad (3.5)$$

as the data which can be correctly classified by $\mathbf{w}^*$ without perturbation. We have

$$\begin{aligned}\nabla_{\mathbf{w}} L_{\mathcal{D}}^{p,r}(\mathbf{w}) &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\ell'(yh(\mathbf{w},\mathbf{x}+\boldsymbol{\delta}))y\nabla_{\mathbf{w}} h(\mathbf{w},\mathbf{x}+\boldsymbol{\delta})\,\mathbb{1}(S)\big] \\ &\quad + \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\ell'(yh(\mathbf{w},\mathbf{x}+\boldsymbol{\delta}))y\nabla_{\mathbf{w}} h(\mathbf{w},\mathbf{x}+\boldsymbol{\delta})\,\mathbb{1}(S^c)\big].\end{aligned}$$

Note $\boldsymbol{\delta} = -ry\bar{\mathbf{w}}/\|\mathbf{w}\|_q^{q-1}$ is the optimal ℓ_p adversarial perturbation corresponding to the model parameter $\mathbf{w}$ and sample $(\mathbf{x},y)$. A routine calculation shows that $yh(\mathbf{w},\mathbf{x}+\boldsymbol{\delta}) = y\mathbf{w}^\top\mathbf{x}/\|\mathbf{w}\|_q - r$. Then it follows that

$$yh(\mathbf{w},\mathbf{x}+\boldsymbol{\delta}) = \begin{cases} \mathrm{sgn}(\mathbf{w}^{*\top}\mathbf{x})\cdot\frac{\mathbf{w}^\top\mathbf{x}}{\|\mathbf{w}\|_q} - r & (\mathbf{x},y)\in S, \\ -\mathrm{sgn}(\mathbf{w}^{*\top}\mathbf{x})\cdot\frac{\mathbf{w}^\top\mathbf{x}}{\|\mathbf{w}\|_q} - r & (\mathbf{x},y)\in S^c, \end{cases}$$

where for the data $(\mathbf{x},y)\in S$ we use $\mathrm{sgn}(\mathbf{w}^{*\top}\mathbf{x})$ to replace the label y while for the data $(\mathbf{x},y)\in S^c$ we use $-\mathrm{sgn}(\mathbf{w}^{*\top}\mathbf{x})$ to replace y. Define

$$\begin{aligned} g_S(\mathbf{w}^*,\mathbf{w};\mathbf{x}) &= \ell'(\mathrm{sgn}(\mathbf{w}^{*\top}\mathbf{x})\cdot\mathbf{w}^\top\mathbf{x}/\|\mathbf{w}\|_q - r), \\ g_{S^c}(\mathbf{w}^*,\mathbf{w};\mathbf{x}) &= \ell'(-\mathrm{sgn}(\mathbf{w}^{*\top}\mathbf{x})\cdot\mathbf{w}^\top\mathbf{x}/\|\mathbf{w}\|_q - r), \\ g(\mathbf{w}^*,\mathbf{w};\mathbf{x}) &= g_S(\mathbf{w}^*,\mathbf{w};\mathbf{x}) + g_{S^c}(\mathbf{w}^*,\mathbf{w};\mathbf{x}). \end{aligned}$$

Then the gradient $\nabla_{\mathbf{w}} L_{\mathcal{D}}^{p,r}(\mathbf{w})$ can be rewritten as

$$\begin{aligned}\nabla_{\mathbf{w}} L_{\mathcal{D}}^{p,r}(\mathbf{w}) &= \mathbb{E}\big[g_S(\mathbf{w}^*,\mathbf{w};\mathbf{x})\mathrm{sgn}(\mathbf{w}^{*\top}\mathbf{x})\nabla_{\mathbf{w}} h(\mathbf{w},\mathbf{x}+\boldsymbol{\delta})\,\mathbb{1}(S)\big] \\ &\quad - \mathbb{E}\big[g_{S^c}(\mathbf{w}^*,\mathbf{w};\mathbf{x})\mathrm{sgn}(\mathbf{w}^{*\top}\mathbf{x})\nabla_{\mathbf{w}} h(\mathbf{w},\mathbf{x}+\boldsymbol{\delta})\,\mathbb{1}(S^c)\big] \\ &= \mathbb{E}\big[g_S(\mathbf{w}^*,\mathbf{w};\mathbf{x})\mathrm{sgn}(\mathbf{w}^{*\top}\mathbf{x})\nabla_{\mathbf{w}} h(\mathbf{w},\mathbf{x}+\boldsymbol{\delta})\big] \\ &\quad - \mathbb{E}\big[g(\mathbf{w}^*,\mathbf{w};\mathbf{x})\mathrm{sgn}(\mathbf{w}^{*\top}\mathbf{x})\nabla_{\mathbf{w}} h(\mathbf{w},\mathbf{x}+\boldsymbol{\delta})\,\mathbb{1}(S^c)\big].\end{aligned}$$

Then using (3.3) and the fact that $\boldsymbol{\delta} = -ry\bar{\mathbf{w}}/\|\mathbf{w}\|_q^{q-1}$, it can be shown that

$$\begin{aligned}\mathbf{w}^{*\top}\nabla L_{\mathcal{D}}^{p,r}(\mathbf{w}) &= \mathbb{E}\big[g_S(\mathbf{w}^*,\mathbf{w};\mathbf{x})\mathrm{sgn}(\mathbf{w}^{*\top}\mathbf{x})\mathbf{w}^{*\top}\nabla_{\mathbf{w}} h(\mathbf{w},\mathbf{x}+\boldsymbol{\delta})\big] \\ &\quad - \mathbb{E}\big[g(\mathbf{w}^*,\mathbf{w};\mathbf{x})\mathrm{sgn}(\mathbf{w}^{*\top}\mathbf{x})\mathbf{w}^{*\top}\nabla_{\mathbf{w}} h(\mathbf{w},\mathbf{x}+\boldsymbol{\delta})\,\mathbb{1}(S^c)\big],\end{aligned} \quad (3.6)$$

where we have defined the quantity $\widetilde{\mathbf{w}} = \mathbf{w}^*/\|\mathbf{w}\|_q - (\bar{\mathbf{w}}^\top\mathbf{w}^*)\mathbf{w}/\|\mathbf{w}\|_q^{q+1}$. This decomposition allows for the label to only play a role through the indicator function $\mathbb{1}(S^c)$.

With this notation in order, we can begin with our proof. The first step is to show that the key quantity (3.6) is more negative when $\mathbf{w}$ is far from $\mathbf{w}^*$ and when the non-robust classification error of the best robust classifier is small.

Lemma 3.3. Let $p\in[2,\infty]$ and $q\in[1,2]$ be such that $1/p + 1/q = 1$ and $\mathbf{w}^* = \arg\min_{\|\mathbf{w}\|_q=1} L_{\mathcal{D}}^{p,r}(\mathbf{w})$ be the optimal model parameter that achieves minimum ℓ_p robust error and $\mathrm{err}_{\mathcal{D}}(\mathbf{w}^*)$ be the clean error achieved by $\mathbf{w}^*$. Suppose the data distribution $\mathcal{D}$ satisfies Assumption

$\nabla L_{\mathcal{D}}^{p,r}(\mathbf{w})$ is defined by $\nabla L_{\mathcal{D}}^{p,r}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\nabla\ell(y\mathbf{w}^\top(\mathbf{x}+\boldsymbol{\delta})/\|\mathbf{w}\|_q)]$, where the gradient is only taken over $\mathbf{w}$ and we do not differentiate through the perturbation $\boldsymbol{\delta}$.

3.1. For any $\mathbf{w}$ of ℓ_q norm 1, let $\widetilde{\mathbf{w}} = \mathbf{w}^* - (\bar{\mathbf{w}}^\top \mathbf{w}^*)\mathbf{w}$, $\theta(\mathbf{w}) = \angle(\mathbf{w}, \mathbf{w}^*)$ and $\theta'(\mathbf{w}) = \angle(-\mathbf{w}, \widetilde{\mathbf{w}})$. Let $\sigma = r$. If $r \le R\|\mathbf{w}\|_2 \sin^{3/2}(\theta'(\mathbf{w}))/(100U)$, $\mathrm{err}_{\mathcal{D}}(\mathbf{w}^*) \le (2^{14}R^4\|\mathbf{w}\|)^{-1}U'^2 r \sin^2(\theta'(\mathbf{w}))$ and

$$\sin(\theta(\mathbf{w})) \ge \max\left\{\frac{4r}{R\|\mathbf{w}\|_2}, \frac{100r\sqrt{U/U'}}{R\|\mathbf{w}\|_2 \sin^{1/2}(\theta'(\mathbf{w}))}\right\},$$

hold, then it holds that

$$\mathbf{w}^{*\top}\nabla L_{\mathcal{D}}^{p,r}(\mathbf{w}) \le -\frac{R^2\|\widetilde{\mathbf{w}}\|_2 \cdot \sin\theta'(\mathbf{w}) \cdot e^{-1}}{2\|\mathbf{w}\|_2}$$

The proof for Lemma 3.3 can be found in Appendix D.1. Lemma 3.3 shows that $\mathbf{w}^{*\top}\nabla L_{\mathcal{D}}^{p,r}(\mathbf{w})$ is more negative when the angle $\theta(\mathbf{w})$ is large, which intuitively means that the distance $\|\mathbf{w} - \mathbf{w}^*\|_2^2$ to the optimal robust classifier will decrease until we reach a point where the angle $\theta(\mathbf{w})$ becomes small (recall the intuition from gradient flow given in (3.4)). We formalize this into the following lemma, which shows that Algorithm 2 leads to a halfspace that is close to the optimal robust classifier.

Lemma 3.4. Let $\delta \in (0,1)$ be arbitrary. Then if $r = O(d^{\frac{3}{2p}-\frac{3}{4}})$, set $\eta = O\big(\delta r^3 d^{\frac{1}{2p}-\frac{1}{4}}\big)$ and run Algorithm $\mathsf{PSAT}(p, r)$ for $K = O\big(d\|\mathbf{w}_1 - \mathbf{w}^*\|_2^2 \delta^{-2} r^{-4} d^{1/2-1/p}\big)$ iterations, with probability at least $1-\delta$, there exists a $k^* \le K$ such that

$$\sin(\theta(\mathbf{w}_{k^*})) \le \begin{cases} O\Big(\frac{r d^{\frac{1}{4}-\frac{1}{2p}}}{\|\mathbf{w}_{k^*}\|_2^{1/2}}\Big) & \|\mathbf{w}_{k^*}\|_2 \ge \|\mathbf{w}^*\|_2, \\ O\Big(\frac{r}{\|\mathbf{w}_{k^*}\|_2}\Big) & \|\mathbf{w}_{k^*}\|_2 < \|\mathbf{w}^*\|_2. \end{cases}$$

The proof for Lemma 3.4 can be found in Appendix D.2. We can now proceed to complete the proof of Theorem 3.2 based on Lemma 3.4 by showing small $\theta(\mathbf{w}_{k^*})$ suffices to ensure small robust classification error $\mathrm{err}_{\mathcal{D}}^{p,r}(\mathbf{w}_{k^*})$. The completed proof of Theorem 3.2 can be found in Appendix D.3 and we sketch the crucial part as follows.

Proof of Theorem 3.2. Before characterizing the robust classification error $\mathrm{err}_{\mathcal{D}}^{p,r}(\mathbf{w}_{k^*})$, we first investigate the optimal robust error $\mathsf{OPT}_{p,r} = \mathrm{err}_{\mathcal{D}}^{p,r}(\mathbf{w}^*)$ and see how it relates to the perturbation radius r. In particular,

$$\begin{aligned}\mathsf{OPT}_{p,r} &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbb{1}\left(y\frac{\mathbf{w}^{*\top}}{\|\mathbf{w}^*\|_q}(\mathbf{x}+\boldsymbol{\delta}) \le 0\right)\right] \\ &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\mathbb{1}(y\mathbf{w}^{*\top}\mathbf{x} \le r)\big],\end{aligned}$$

where we use the fact that $\|\mathbf{w}^*\|_q = 1$ in the second equality. Note that the robust error consists of two disjoint parts of data: (1) the data satisfies $|\mathbf{w}^{*\top}\mathbf{x}| \le r$; and (2) the data satisfies $|\mathbf{w}^{*\top}\mathbf{x}| > r$ and $y\mathbf{w}^{*\top}\mathbf{x} < 0$. Therefore, we can get lower and upper bounds on $\mathsf{OPT}_{p,r}$,

$$\begin{aligned}&\mathsf{OPT}_{p,r} \ge \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_x}\big[\mathbb{1}(|\mathbf{w}^{*\top}\mathbf{x}| \le r)\big] \\ &\mathsf{OPT}_{p,r} \le \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_x}\big[\mathbb{1}(|\mathbf{w}^{*\top}\mathbf{x}| \le r)\big] + \mathrm{err}_{\mathcal{D}}(\mathbf{w}^*). \quad (3.7)\end{aligned}$$

By Assumption 3.1, we have the data distribution $\mathcal{D}_x$ satisfies U-anti-concentration and (U', R) anti-anti-concentration with U, R being constants. Therefore, it follows that $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_x}\big[\mathbb{1}(|\mathbf{w}^{*\top}\mathbf{x}| \le r)\big] = \Theta(r\|\mathbf{w}^*\|_2^{-1})$ since we have $r\|\mathbf{w}\|_2^{-1} = O(d^{\frac{3}{2p}-\frac{3}{4}}\|\mathbf{w}\|_2^{-1}) \le R$. Besides, note that we also have $\mathrm{err}_{\mathcal{D}}(\mathbf{w}^*) = O(rd^{2/p-1}) \le O(r\|\mathbf{w}^*\|_2^{-1})$ due to our assumption. Therefore, it is clear that $\mathsf{OPT}_{p,r} = \Theta(r\|\mathbf{w}^*\|_2^{-1})$.

An argument similar to that used for (3.7) leads to the bound

$$\begin{aligned}\mathrm{err}_{\mathcal{D}}^{p,r}(\mathbf{w}_{k^*}) &\le \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_x}\big[\mathbb{1}(|\mathbf{w}_{k^*}^\top\mathbf{x}| \le r)\big] + \mathrm{err}_{\mathcal{D}}(\mathbf{w}^*) \\ &= O(r\|\mathbf{w}_{k^*}\|_2^{-1}) + \mathrm{err}_{\mathcal{D}}(\mathbf{w}_{k^*}). \quad (3.8)\end{aligned}$$

We proceed by sketching how we bound each of these two terms. For $O(r\|\mathbf{w}_{k^*}\|_2^{-1})$, we only need to characterize the ℓ_2 norm of $\mathbf{w}_{k^*}$. In fact by Lemma 3.4, we can show that under the assumption that $r = O(d^{\frac{3}{2p}-\frac{3}{4}})$ it holds that $\|\mathbf{w}_{k^*}\|_2 = \Omega(\|\mathbf{w}^*\|_2)$ (see Appendix D.3 for more details), which further implies that $O(r\|\mathbf{w}_{k^*}\|_2^{-1}) = O(r\|\mathbf{w}^*\|_2^{-1}) = O(\mathsf{OPT}_{p,r})$.

The next step is to characterize $\mathrm{err}_{\mathcal{D}}(\mathbf{w}_{k^*})$. We can do so by comparing it with the classification error of $\mathbf{w}^*$,

$$\begin{aligned}\mathrm{err}_{\mathcal{D}}(\mathbf{w}_{k^*}) &\le \mathrm{err}_{\mathcal{D}}(\mathbf{w}^*) + |\mathrm{err}_{\mathcal{D}}(\mathbf{w}_{k^*}) - \mathrm{err}_{\mathcal{D}}(\mathbf{w}^*)| \\ &\le 2\mathrm{err}_{\mathcal{D}}(\mathbf{w}^*) + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_x}[\mathbb{1}(\mathbf{w}_{k^*}^\top\mathbf{x} \ne \mathbf{w}^{*\top}\mathbf{x})] \\ &\le O(r\|\mathbf{w}^*\|_2^{-1}) + \Theta(\theta(\mathbf{w}_{k^*})), \quad (3.9)\end{aligned}$$

where the last inequality is due to the fact that $\mathcal{D}_x$ is isotropic (see Appendix D.3 for more details). Then by Lemma 3.4 it is clear that

$$\theta(\mathbf{w}_{k^*}) = O\left(\frac{r d^{\frac{1}{4}-\frac{1}{2p}}}{\|\mathbf{w}_{k^*}\|_2^{1/2}}\right) = O\left(\frac{r d^{\frac{1}{4}-\frac{1}{2p}}}{\|\mathbf{w}^*\|_2^{1/2}}\right) \quad (3.10)$$

since we have shown that $\|\mathbf{w}_{k^*}\|_2 = \Omega(\|\mathbf{w}^*\|_2)$. Consequently, combining (3.10) and (3.9) and further substituting into (3.8), we get that $\mathrm{err}_{\mathcal{D}}^{p,r}(\mathbf{w}_{k^*})$ is at most

$$O\left(\frac{r d^{\frac{1}{4}-\frac{1}{2p}}}{\|\mathbf{w}^*\|_2^{1/2}}\right) = O\big(d^{\frac{1}{4}-\frac{1}{2p}} \cdot \|\mathbf{w}^*\|_2^{1/2} \cdot \mathsf{OPT}_{p,r}\big)$$

since $\mathsf{OPT}_{p,r} = \Theta(r\|\mathbf{w}^*\|_2^{-1})$. This completes the proof. □

4. Conclusion and Future Work

In this work we analyzed the properties of adversarial training for learning halfspaces with noise. We provided the first guarantee that adversarial training provably leads to robust classifiers when the data distribution has label noise. In particular, we established that adversarial training leads to approximate minimizers for the robust classification error

under ℓ^p perturbations for many distributions. For typical convex loss functions like the cross entropy or hinge loss, we showed that adversarial training can achieve robust classification error $\widetilde{O}\big(\sqrt{\mathsf{OPT}_{2,r}}\big)$ when $p = 2$ and $\widetilde{O}\big(d^{1/4}\sqrt{\mathsf{OPT}_{\infty,r}} + d^{1/2}\mathsf{OPT}_{\infty,r}\big)$ for ℓ_∞ when $p = \infty$ for distributions satisfying anti-concentration properties. We showed that the robust classification error guarantees can be improved if we instead use a nonconvex sigmoidal loss, with guarantees of $O(\mathsf{OPT}_{2,r})$ for $p = 2$ and $O(d^{1/4}\mathsf{OPT}_{\infty,r})$ for $p = \infty$ in the worst case. For future work, we are keen on understanding whether or not adversarial training provably leads to robust classifiers for more complicated function classes than halfspaces.

Acknowledgement

We would like to thank the anonymous reviewers for their helpful comments. DZ is supported by the Bloomberg Data Science Ph.D. Fellowship. SF is supported by the UCLA Dissertation Year Fellowship. QG is partially supported by the National Science Foundation IIS-2008981. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Balcan, M.-F. F. and Zhang, H. Sample and computationally efficient learning algorithms under s-concave distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, October 2009.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Conference on Learning Theory (COLT)*, 1992.

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

Charles, Z., Rajput, S., Wright, S., and Papailiopoulos, D. Convergence and margin of adversarial training on separable data. *Preprint, arXiv:1905.09209*, 2019.

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2019.

Cullina, D., Bhagoji, A. N., and Mittal, P. Pac-learning in the presence of evasion adversaries. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Daniely, A. Complexity theoretic limitations on learning halfspaces. In *ACM Symposium on Theory of Computing (STOC)*, pp. 105–117, 2016.

Diakonikolas, I., Kane, D. M., and Manurangsi, P. The complexity of adversarially robust proper learning of halfspaces with agnostic noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020a.

Diakonikolas, I., Kontonis, V., Tzamos, C., and Zarifis, N. Learning halfspaces with massart noise under structured distributions. In *Conference on Learning Theory (COLT)*, 2020b.

Diakonikolas, I., Kontonis, V., Tzamos, C., and Zarifis, N. Non-convex sgd learns halfspaces with adversarial label noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020c.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pp. 272–279, 2008.

Frei, S., Cao, Y., and Gu, Q. Agnostic learning of halfspaces with gradient descent via soft margins. *arXiv preprint: 2010.00539*, 2020.

Gao, R., Cai, T., Li, H., Wang, L., Hsieh, C.-J., and Lee, J. D. Convergence of adversarial training in overparametrized neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. URL http://arxiv.org/abs/1412.6572.

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 (NeurIPS)*, 2019.

Javanmard, A., Soltanolkotabi, M., and Hassani, H. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory (COLT)*, 2020.

Kakade, S. M., Sridharan, K., and Tewari, A. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2009.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. *Preprint, arXiv:1611.01236*, 2016.

Lécuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with

differential privacy. In *IEEE Symposium on Security and Privacy, SP 2019*, pp. 656–672, 2019.

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

Lovász, L. and Vempala, S. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Montasser, O., Goel, S., Diakonikolas, I., and Srebro, N. Efficiently learning adversarially robust halfspaces with noise. In *International Conference on Machine Learning (ICML)*, 2020.

Nandy, J., Saha, S., Hsu, W., Lee, M. L., and Zhu, X. X. Adversarially trained models with test-time covariate shift adaptation. *Preprint, arXiv:2102.05096*, 2021.

Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 582–597, 2016. doi: 10.1109/SP.2016.41.

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J., and Liang, P. Understanding and mitigating the tradeoff between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2020.

Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Salman, H., Li, J., Razenshteyn, I., Zhang, P., Zhang, H., Bubeck, S., and Yang, G. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., and Madry, A. Do adversarially robust imagenet models transfer better? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. URL http://arxiv.org/abs/1312.6199.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, 2019.

Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., and Gu, Q. On the convergence and robustness of adversarial training. In *ICML*, volume 1, pp. 2, 2019a.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., and Gu, Q. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019b.

Wong, E. and Kolter, J. Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning (ICML)*, 2018.

Wu, B., Chen, J., Cai, D., He, X., and Gu, Q. Does network width really help adversarial robustness? *arXiv preprint arXiv:2010.01279*, 2020.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R., and Chaudhuri, K. A closer look at accuracy vs. robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019.

Zhang, Y., Plevrakis, O., Du, S. S., Li, X., Song, Z., and Arora, S. Over-parameterized adversarial training: An analysis overcoming the curse of dimensionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.