Revisiting Adversarial Risk

Arun Sai Suggala **Adarsh Prasad** **Vaishnavh Nagarajan** **Pradeep Ravikumar**
Carnegie Mellon University
{asuggala, adarshp, vaishnavh, pradeepr}@cs.cmu.edu

Abstract

Recent works on adversarial perturbations show that there is an inherent trade-off between standard test accuracy and adversarial accuracy. Specifically, they show that no classifier can simultaneously be robust to adversarial perturbations and achieve high standard test accuracy. However, this is contrary to the standard notion that on tasks such as image classification, humans are robust classifiers with low error rate. In this work, we show that the main reason behind this confusion is the inexact definition of adversarial perturbation that is used in the literature. To fix this issue, we propose a slight, yet important modification to the existing definition of adversarial perturbation. Based on the modified definition, we show that there is no trade-off between adversarial and standard accuracies; there exist classifiers that are robust and achieve high standard accuracy. We further study several properties of this new definition of adversarial risk and its relation to the existing definition.

1 Introduction

Recent works have shown that the output of deep neural networks is vulnerable to even a small amount of perturbation to the input [Goodfellow et al., 2014, Szegedy et al., 2013]. These perturbations, usually referred to as "adversarial" perturbations, are imperceivable by humans and can deceive even state-of-the-art models to make incorrect predictions. Consequently, a line of work in deep learning has focused on defending against such attacks/perturbations [Goodfellow et al., 2014, Carlini and Wagner, 2016, Ilyas et al., 2017, Madry et al., 2017]. This has resulted in several techniques for learning models that are robust to adversarial attacks. However, many of these techniques were later shown to be ineffective [Athalye and Sutskever, 2017, Carlini and Wagner, 2017, Athalye et al., 2018].

We present a brief review of existing literature on adversarial robustness, that is necessarily incomplete. Existing works define an adversarial perturbation at a point $\mathbf{x}$, for a classifier f as any perturbation $\boldsymbol{\delta}$ with a small norm, measured w.r.t some distance metric, which changes the output of the classifier; that is $f(\mathbf{x}+\boldsymbol{\delta}) \neq f(\mathbf{x})$. Most of the existing techniques for learning robust models minimize the following worst case loss over all possible perturbations

$$\mathbb{E}_{(\mathbf{x},y)\sim P}\left[\max_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|\leq\epsilon} \ell(f(\mathbf{x}+\boldsymbol{\delta}), y)\right]. \qquad (1)$$

Goodfellow et al. [2014], Carlini and Wagner [2017], Madry et al. [2017] use heuristics to approximately minimize the above objective. In each iteration of the optimization, these techniques first use heuristics to approximately solve the inner maximization problem and then compute a descent direction using the resulting maximizers. Tsuzuku et al. [2018] provide a training algorithm which tries to find large margin classifiers with small Lipschitz constants, thus ensuring robustness to adversarial perturbations. A recent line of work has focused on optimizing an upper bound of the above objective. Raghunathan et al. [2018], Kolter and Wong [2017] provide SDP and LP based upper bound relaxations of the objective, which can be solved efficiently for small networks. These techniques have the added advantage that they can be used to formally verify the robustness of any given model. Sinha et al. [2017] propose to optimize the following distributional robustness objective, which is a stronger form of robustness than the one used in Equation (1)

$$\min_{f} \sup_{Q:W(P,Q)\leq\epsilon} \mathbb{E}_{(\mathbf{x},y)\sim Q}\left[\ell(f(\mathbf{x}), y)\right], \qquad (2)$$

where $W(P,Q)$ is the Wasserstein distance between probability distributions P, Q.

Another line of work on adversarial robustness has focused on studying adversarial risk from a theoretical

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

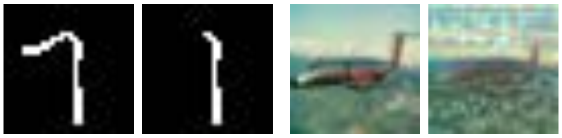
perspective. Recently, Schmidt et al. [2018], Bubeck et al. [2018] study the generalization properties of adversarial risk and compare it with the generalization properties of standard risk ($\mathbb{P}(y \neq f(\mathbf{x}))$). Fawzi et al. [2018], Fawzi et al. [2018], Franceschi et al. [2018] study the properties of adversarial perturbations and adversarial risk. These works characterize the robustness at a point $\mathbf{x}$ in terms of how much perturbation a classifier can tolerate at a point, without changing its prediction

$$r(\mathbf{x}) = \min_{\boldsymbol{\delta}\in\mathcal{S}} \|\boldsymbol{\delta}\| \quad s.t.\ \mathrm{sign}(f(\mathbf{x})) \neq \mathrm{sign}(f(\mathbf{x}+\boldsymbol{\delta})), \quad (3)$$

where $\mathcal{S}$ is some subspace. Fawzi et al. [2018] theoretically study the expected adversarial radius ($\mathbb{E}[r(\mathbf{x})]$) of any classifier f and suggest that there is a trade-off between adversarial robustness and the standard accuracy. Specifically, their results suggest that if the prediction accuracy is high then $\mathbb{E}[r(\mathbf{x})]$ could be small.

However, these results are contrary to the standard notion that on tasks such as image classification, humans are robust classifiers with low error rate. A careful inspection of the definition of adversarial perturbation and adversarial radius used in Equations (1),(3) brings into light the inaccuracies of these definitions. For example, consider the definition of adversarial risk in Equation (1). A major issue with this definition is that it assumes the label y remains the same in a neighborhood of $\mathbf{x}$, and penalizes any classifier which doesn't output y in the entire neighborhood of $\mathbf{x}$. However, the response variable need not remain the same in the neighborhood of $\mathbf{x}$. If a perturbation $\boldsymbol{\delta}$ is such that "true label" at $\mathbf{x}$ is not the same as the "true label" at $\mathbf{x}+\boldsymbol{\delta}$ then the classifier shouldn't be penalized for not predicting y at $\mathbf{x}+\boldsymbol{\delta}$. Moreover, such a perturbation shouldn't be considered as adversarial, since it changes the true label at $\mathbf{x}+\boldsymbol{\delta}$. Figure 1 illustrates this phenomenon on MNIST and CIFAR-10. As we show later in the paper, this inexact definition of adversarial perturbation has resulted in recent works claiming that there exists a trade-off between adversarial and standard risks.

To be more concrete, consider two points $(\mathbf{x}, 1)$ and $(\mathbf{x}+\boldsymbol{\delta}, -1)$ which are close to each other ($i.e.$, $\|\boldsymbol{\delta}\| \leq \epsilon$). Then for any classifier to be correct at the two points, it has to change its prediction for the two points over a small region, which would mean that the adversarial radius, $r(\mathbf{x})$, is very small. This shows that in order to have high accuracy, a classifier will have to change its score over a small region, leading to a small adversarial radius. This creates the illusion of a trade-off between adversarial robustness and standard risk. This illusion arises because of the above definitions of adversarial perturbation which consider the perturbation $\boldsymbol{\delta}$ at $\mathbf{x}$ to be adversarial. On the contrary, $\boldsymbol{\delta}$ shouldn't be considered adversarial because the true label at $\mathbf{x}+\boldsymbol{\delta}$ is not the same as the label at $\mathbf{x}$. This confusion motivates the need for a clear definition of an adversarial perturbation, the corresponding adversarial risk, and then studying these quantities.



(a) MNIST (L_0) (b) CIFAR-10 (L_2)

Figure 1: Images from Sharif et al. [2018] showing that small adversarial perturbations can change the true label of the perturbed image. Left and right images in each sub-figure correspond to the original and perturbed images. 4.5% of the pixels are corrupted by the L_0 adversary and $\epsilon = 6$ for the L_2 adversary.

Contributions. In this work, we first formally define the notions of adversarial perturbation, adversarial risk, which address the above described issue with the existing definition of adversarial risk. Next, we present two key sets of results. One set of results pertain to our modified definition of adversarial risk (Sections 4, 6). In Section 4 we show that the minimizers of both adversarial and standard training objectives are Bayes optimal classifiers. This shows that there is no trade-off between adversarial and standard risks and there exist classifiers which have low adversarial and standard risks. Despite this result, in Section 6, we show that there is a need for adversarial training. The second set of results in Section 5 analyze the existing definition of adversarial risk to answer some natural questions that come up in light of our results in Section 4. Specifically, we study the conditions under which similar results as in Section 4 hold for the existing definition of adversarial risk.

2 Preliminaries

In this section, we set up the notation and review necessary background on risk minimization. To simplify the presentation in the paper, we only consider the binary classification problem. However, it is straightforward to extend the results and analysis in this paper to multi-class classification.

Let $(\mathbf{x}, y) \in \mathbb{R}^d \times \{-1, 1\}$ denote the covariate, label pair which follows a probability distribution P. Let $S_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be n i.i.d samples drawn from P. Let $f : \mathbb{R}^d \to \mathbb{R}$ denote a score based classifier, which assigns $\mathbf{x}$ to class 1, if $f(\mathbf{x}) > 0$. We define the population and empirical risks of classifier f as

$$R_{0-1}(f) = \mathbb{E}_{(\mathbf{x},y)\sim P}\left[\ell_{0-1}(f(\mathbf{x}), y)\right],$$
$$R_{n,0-1}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell_{0-1}(f(\mathbf{x}_i), y_i),$$

where $\ell_{0-1}(\cdot,\cdot)$ is defined as $\ell_{0-1}(f(\mathbf{x}),y) = \mathbb{I}(\text{sign}(f(\mathbf{x})) \neq y)$, and $\text{sign}(\alpha) = 1$ if $\alpha > 0$ and -1 otherwise. Given S_n, the objective of empirical risk minimization (ERM) is to estimate a classifier with low population risk $R(f)$. Since optimization of 0/1 loss is computationally intractable, it is often replaced with a convex surrogate loss function $\ell(f(\mathbf{x}),y) = \phi(yf(\mathbf{x}))$, where $\phi : \mathbb{R} \to [0,\infty)$. Logistic loss is a popularly used surrogate loss and is defined as $\ell(f(\mathbf{x}),y) = \log(1 + e^{-yf(\mathbf{x})})$. We let $R(f), R_n(f)$ denote the population and empirical risk functions obtained by replacing ℓ_{0-1} with ℓ in $R_{0-1}(f), R_{n,0-1}(f)$.

A score based classifier f^* is called Bayes optimal classifier if $\text{sign}(f^*(\mathbf{x})) = \text{sign}(2P(y=1|\mathbf{x})-1)$ a.e. on the support of distribution P. We call $\eta(\mathbf{x}) = \text{sign}(f^*(\mathbf{x}))$ as Bayes decision rule. Note that Bayes decision rule need not be unique. There could be multiple Bayes decision rules which differ on points outside the support of P. We assume that the set of points where $P(y=1|\mathbf{x}) = \frac{1}{2}$ has measure 0.

3 Adversarial Risk

In this paper, we focus on the following robustness setting, which is also the focus of most of the past works on adversarial robustness: given a pre-trained model, there is an adversary which corrupts the inputs to the model such that the corrupted inputs lead to certain "unwanted" behavior in the model. Our goal is to design models that are robust to such adversaries. In what follows, we make the notions of an adversary, unwanted behavior more concrete and formally define adversarial perturbation and adversarial risk.

Let $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^d$ be an adversary which modifies any given data point $\mathbf{x}$ to $\mathcal{A}(\mathbf{x})$. Let $\boldsymbol{\delta}_\mathbf{x} = \mathcal{A}(\mathbf{x}) - \mathbf{x}$ be the perturbation chosen by the adversary at $\mathbf{x}$. We assume that the perturbations are norm bounded, which is a standard restriction imposed on the capability of the adversary.

Our definition of adversarial perturbation is based on a reference or a base classifier. For example, in vision tasks, this base classifier is the human vision system. A perturbation is adversarial to a classifier if it changes the prediction of the classifier, whereas the base/reference classifier assigns it to the same class as the unperturbed point.

Definition 1 (Adversarial Perturbation). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a score based classifier and $g : \mathbb{R}^d \to \{-1,1\}$ be a base classifier. Then the perturbation $\boldsymbol{\delta}_\mathbf{x}$ chosen by an adversary $\mathcal{A}$ at $\mathbf{x}$ is said to be adversarial for f, w.r.t base classifier g, if $\|\boldsymbol{\delta}_\mathbf{x}\| \leq \epsilon$ and*

$$sign(f(\mathbf{x})) = g(\mathbf{x}), \quad g(\mathbf{x}) = g(\mathbf{x} + \boldsymbol{\delta}_\mathbf{x}),$$

and

$$sign(f(\mathbf{x} + \boldsymbol{\delta}_\mathbf{x})) \neq g(\mathbf{x}).$$

Equivalently, a perturbation $\boldsymbol{\delta}_\mathbf{x}$ is said to be adversarial for f, w.r.t base classifier g, if $\|\boldsymbol{\delta}_\mathbf{x}\| \leq \epsilon$, $g(\mathbf{x}) = g(\mathbf{x} + \boldsymbol{\delta}_\mathbf{x})$ and

$$\ell_{0-1}\left(f(\mathbf{x} + \boldsymbol{\delta}_\mathbf{x}), g(\mathbf{x})\right) - \ell_{0-1}\left(f(\mathbf{x}), g(\mathbf{x})\right) = 1.$$

Note that, unlike the existing notion of adversarial risk, the above definition doesn't consider a perturbation as adversarial if it changes the label of the base classifier. Moreover, if f disagrees with g at $\mathbf{x}$, then the perturbation $\boldsymbol{\delta}_\mathbf{x}$ is not considered adversarial. This is reasonable because if $f(\mathbf{x})$ disagrees with $g(\mathbf{x})$, it should be treated as a standard classification error rather than adversarial error. Using the above definition of adversarial perturbation, we next define adversarial risk.

Definition 2 (Adversarial Risk). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a score based classifier and $g : \mathbb{R}^d \to \{-1,1\}$ be a base classifier. The adversarial risk of f w.r.t base classifier g and adversary $\mathcal{A}$ is defined as the fraction of points which can be adversarially perturbed by $\mathcal{A}$*

$$R_{adv,0-1}(f) = \mathbb{E}\left[\ell_{0-1}\left(f(\mathbf{x} + \boldsymbol{\delta}_\mathbf{x}), g(\mathbf{x})\right) - \ell_{0-1}\left(f(\mathbf{x}), g(\mathbf{x})\right)\right].$$

It is typically assumed that the adversary $\mathcal{A}$ is an "optimal" adversary; that is, at any give point $\mathbf{x}$, $\mathcal{A}$ tries to find a perturbation that is adversarial for f

$$\boldsymbol{\delta}_\mathbf{x} \in \underset{\substack{\|\boldsymbol{\delta}\| \leq \epsilon \\ g(\mathbf{x}) = g(\mathbf{x}+\boldsymbol{\delta})}}{\text{argmax}} \; \ell_{0-1}\left(f(\mathbf{x} + \boldsymbol{\delta}), g(\mathbf{x})\right) - \ell_{0-1}\left(f(\mathbf{x}), g(\mathbf{x})\right).$$

The adversarial risk of a classifier f w.r.t an optimal adversary can then be written as

$$\mathbb{E}\left[\max_{\substack{\|\boldsymbol{\delta}\| \leq \epsilon \\ g(\mathbf{x}) = g(\mathbf{x}+\boldsymbol{\delta})}} \ell_{0-1}\left(f(\mathbf{x} + \boldsymbol{\delta}), g(\mathbf{x})\right) - \ell_{0-1}\left(f(\mathbf{x}), g(\mathbf{x})\right)\right].$$

In the sequel, we assume that the adversary is optimal and work with the above definition of adversarial risk.

Let $R_{\text{adv}}(f)$ denote the adversarial risk obtained by replacing ℓ_{0-1} with a convex surrogate loss ℓ and let $R_{n,\text{adv}}(f)$ denote its empirical version. In the sequel we refer to $R(f), R_{\text{adv}}(f)$ as standard and adversarial risks and $R_n(f)$, $R_{n,\text{adv}}(f)$ as the corresponding empirical risks. The goal of adversarial training is to learn a classifier that has low adversarial and standard risks. One natural technique to estimate such a robust classifier is to minimize a linear combination of both the risks

$$\underset{f \in \mathcal{F}}{\text{argmin}}\, R(f) + \lambda R_{\text{adv}}(f), \tag{4}$$

where $\mathcal{F}$ is an appropriately chosen function class and $\lambda \geq 0$ is a hyper-parameter. The tuning parameter λ trades off standard risk with the *excess risk* incurred from adversarial perturbations, and allows us to tune the conservativeness of our classifier.

4 Bayes Optimal Classifier as Base Classifier

In this section we study the properties of minimizers of objective (4), under the assumption that the base classifier $g(\mathbf{x})$ is a Bayes optimal classifier. This is a reasonable assumption because if we are interested in robustness with respect to a base classifier, it is likely we are getting labels from the base classifier itself. For instance, in many classification tasks the labels are generated by humans (*i.e.,* human is a Bayes optimal classifier for the classification task) and robustness is also measured w.r.t a human. The following Theorem shows that under this condition, the minimizers of (4) are Bayes optimal.

Theorem 1. *Suppose the hypothesis class $\mathcal{F}$ is the set of all measurable functions. Let the base classifier $g(x)$ be a Bayes optimal classifier.*

1. *(0/1 **loss**). If ℓ is the 0/1 loss, then any minimizer $\hat{f}$ of*
$$\min_{f\in\mathcal{F}} R_{0-1}(f) + \lambda R_{adv,0-1}(f),$$
is a Bayes optimal classifier.
2. *(**Logistic loss**). Suppose ℓ is the logistic loss and suppose the probability distribution P is such that $\left|P(y=1|\mathbf{x}) - \frac{1}{2}\right| > \gamma$ a.e., for some positive constant γ. Then any minimizer of Equation (4) is a Bayes optimal classifier.*

The first part of the above Theorem shows that minimizing the joint objective with 0/1 loss, for any choice of $\lambda \geq 0$, results in a Bayes optimal classifier. This shows that there exist classifiers that are both robust and achieve high standard accuracy and there is no trade-off between adversarial and standard risks. More importantly, the Theorem shows that if there exists a unique Bayes decision rule (*i.e.,* $\mathrm{sign}(f_1^*) = \mathrm{sign}(f_2^*)$ a.e. for any two Bayes optimal classifiers f_1, f_2), then standard training suffices to learn robust classifiers and there is no need for adversarial training.

The second part of the Theorem, which is perhaps the more interesting result, shows that using a convex surrogate for the 0/1 loss to minimize the joint objective also results in Bayes optimal classifiers. This result assures us that optimizing a convex surrogate does not hinder our search for a robust classifier that has low adversarial and standard risks. Finally, we note that the requirement on conditional class probability $P(y|\mathbf{x})$ is a mild condition as γ can be any small positive constant close to 0.

4.1 Approximate Bayes Optimal Classifier as Base Classifier

We now briefly discuss the scenario where the base classifier $g(\mathbf{x})$ is not Bayes optimal. In this setting, the minimizers of the objective (4) need not be Bayes optimal. The first term in the objective will bias the optimization towards a Bayes optimal classifier. Whereas, the second term in the joint objective will bias the optimization towards the base classifier. Since the base classifier is not a Bayes optimal classifier, this results in a trade-off between the two terms, which is controlled by the tuning parameter λ. If λ is small, then the minimizers of the joint objective will be close to a Bayes optimal classifier. If λ is large, the minimizers will be close to the base classifier.

5 Old definition of Adversarial Risk

One natural question that Section 4 gives rise to is whether the results in Theorem 1 also hold for the definition of adversarial risk used by the existing works. To answer this question, we now study the properties of minimizers of the adversarial training objective in Equation (1). We start by making a slight modification to the definition of adversairal risk $R_{\mathrm{adv},0-1}(f)$ and analyzing the minimizers of the resulting adversarial training objective. Let $H_{\mathrm{adv},0-1}(f)$ be the adversarial risk obtained by removing the constraint $g(\mathbf{x}+\boldsymbol{\delta}) = g(\mathbf{x})$ in $R_{\mathrm{adv},0-1}(f)$

$$H_{\mathrm{adv},0-1}(f) = \mathbb{E}\left[\sup_{\|\boldsymbol{\delta}\|\leq\epsilon} \ell_{0-1}\left(f(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x})\right) - \ell_{0-1}\left(f(\mathbf{x}), g(\mathbf{x})\right)\right].$$

We call this the adversarial "smooth" risk, because by removing the constraint, we are implicitly assuming that the base classifier is smooth in the neighborhood of each point. Let $H_{\mathrm{adv}}(f)$ denote the adversarial risk obtained by replacing ℓ_{0-1} in $H_{\mathrm{adv},0-1}(f)$ with a convex surrogate loss ℓ.

The following Theorem studies the minimizers of the adversarial training objective obtained using the adversarial smooth risk. Specifically, it shows that if there exists a Bayes decision rule which satisfies a "margin condition", then minimizing the adversarial training objective using $H_{\mathrm{adv},0-1}(f)$ results in Bayes optimal classifiers.

Theorem 2. *Suppose the hypothesis class $\mathcal{F}$ is the set of all measurable functions. Moreover, suppose there exists a Bayes decision rule $\eta(\mathbf{x})$ which satisfies the following margin condition:*

$$Pr\left(\{\mathbf{x} : \exists\tilde{\mathbf{x}}, \|\tilde{\mathbf{x}}-\mathbf{x}\| \leq \epsilon \text{ and } \eta(\tilde{\mathbf{x}}) \neq \eta(\mathbf{x})\}\right) = 0. \quad (5)$$

1. *(0/1 **loss**). If ℓ is the 0/1 loss, then any minimizer of $R_{0-1}(f) + \lambda H_{adv,0-1}(f)$ is a Bayes optimal classifier.*
2. *(**Logistic loss**). Suppose ℓ is the logistic loss. Moreover, suppose the probability distribution P is such that $\left|P(y=1|\mathbf{x}) - \frac{1}{2}\right| > \gamma$ a.e., for some positive constant γ. Then any minimizer of*
$$\min_{f\in\mathcal{F}} R(f) + \lambda H_{adv}(f), \quad (6)$$

is a Bayes optimal classifier

The margin condition in Equation (5) requires the Bayes decision rule to *not* change its prediction in the neighborhood of any given point. We note that this condition is necessary for the results of the above Theorem to hold. In Section 5.2 we show that without the margin condition, the minimizers of (6) need not be Bayes optimal. Theorem 2 also highlights the importance of the constraint "$g(\mathbf{x}+\boldsymbol{\delta}) = g(\mathbf{x})$" in the definition of adversarial risk, for Bayes optimality of the minimizers.

5.1 Replacing Base Classifier with Stochastic Label y

We now proceed to study the properties of minimizers of (1). We replace $g(\mathbf{x})$ in the definition of adversarial smooth risk $H_{\text{adv},0-1}(\cdot)$ with stochastic label y and study the properties of minimizers of the resulting objective. Our results show that the resulting adversarial training objective behaves similarly as Equation (6).

Theorem 3. *Consider the setting of Theorem 2. Let $G_{adv,0-1}(f)$ be the adversarial risk obtained by replacing $g(\mathbf{x})$ with y in $R_{adv,0-1}(f)$*

$$G_{adv,0-1}(f) = \mathbb{E}\left[\sup_{\|\boldsymbol{\delta}\|\leq\epsilon} \ell_{0-1}\left(f(\mathbf{x}+\boldsymbol{\delta}), y\right) - \ell_{0-1}\left(f(\mathbf{x}), y\right)\right].$$

1. *(0/1 **loss**). If ℓ is the 0/1 loss, then any minimizer of $R_{0-1}(f) + \lambda G_{adv,0-1}(f)$ is a Bayes optimal classifier.*
2. *(**Logistic loss**). Suppose ℓ is the logistic loss. Moreover, suppose the probability distribution P is such that $\left|P(y=1|\mathbf{x}) - \frac{1}{2}\right| > \gamma$ a.e., for some positive constant γ. Then any minimizer of*

$$\min_{f\in\mathcal{F}} R(f) + \lambda G_{adv}(f), \tag{7}$$

is a Bayes optimal classifier

Note that, objective (1) is equivalent to objective (7) for $\lambda = 1$. The Theorem thus shows that under the margin condition there is no trade-off between the popularly used definition of adversarial risk and standard risk.

5.2 Importance of Margin

If no Bayes decision rule satisfies the margin condition, then the results of Theorems 2,3 do not hold and minimizers of the corresponding joint objectives need not be Bayes optimal.

Theorem 4 (Necessity of margin)**.** *Consider the setting of Theorem 2. Suppose no Bayes decision rule satisfies the margin condition in Equation (5). Then $\exists\lambda_0$ such that $\forall\lambda > \lambda_0$ the minimizers of the joint objectives $R_{0-1}(f) + \lambda H_{adv,0-1}(f)$ and $R_{0-1}(f) + \lambda G_{adv,0-1}(f)$ are not Bayes optimal.*

The above Theorem shows that without the margin condition, performing adversarial training using existing definition of adversarial risk can result in a loss of standard accuracy. Next, we consider a concrete example and empirically validate our findings from Theorems 3, 4.

Synthetic Dataset. Consider the following data generation process in a 2D space. Let $S(\mathbf{c}, r)$ denote the axis aligned square of side length r, centered at $\mathbf{c}$. The marginal distribution of $\mathbf{x}$ follows a uniform distribution on $S([-2,0]^T, 2) \cup S([2,0]^T, 2)$. The conditional distribution of y given $\mathbf{x}$ is given by

$$y|\mathbf{x} \in S([2,0]^T, 2) = \begin{cases} 1, & \text{w.p. } 0.7 \\ -1, & \text{w.p. } 0.3 \end{cases},$$

$$y|\mathbf{x} \in S([-2,0]^T, 2) = \begin{cases} 1, & \text{w.p. } 0.3 \\ -1, & \text{w.p. } 0.7 \end{cases}.$$

Note that the data satisfies the margin condition in Equation (5) w.r.t L_∞ norm, for $\epsilon = 1$ and the following Bayes decision rule

$$\eta(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x}(1) \geq 0 \\ -1 & \text{if } \mathbf{x}(1) < 0 \end{cases}.$$

From Theorem 3 we know that for L_∞ norm perturbations with $\epsilon \leq 1$, minimizing Equation 7 results in Bayes optimal classifiers. To verify this, we generated 10^5 training samples from this distribution and minimized objective (7) over the set of linear classifiers. Since the model is linear, we have a closed form expression for the adversarial risk. Moreover, objective (7) can be efficiently solved using gradient descent. Figure 2 shows the behavior of standard risk of the resulting models as we vary ϵ. We can seen that for $\epsilon \leq 1$, the standard risk is equal to 0.3, which is the Bayes optimal risk. Whereas, for $\epsilon > 1$, the standard risk can be larger than 0.3.

Benchmark Datasets. A number of recent works try to explain the drop in standard accuracy in adversarially trained models [Fawzi et al., 2018, Tsipras et al., 2018]. These works suggest that there could be an inherent trade-off between standard and adversarial risks. In contrast, our results show that as long as there exists a Bayes optimal classifier with sufficient margin, minimizers of objectives (1), (7) have low standard and adversarial risks and there is no trade-off between the two risks. The important question then is, "Do the benchmark datasets such as MNIST [LeCun, 1998], CIFAR10 [Krizhevsky and Hinton, 2009] satisfy

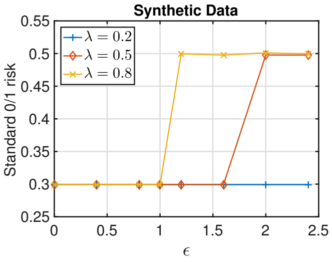


Figure 2: Figure shows standard 0/1 risk vs. ϵ on the synthetic dataset. The adversarial perturbations are measured w.r.t L_∞ norm.

the margin condition?" Sharif et al. [2018] try to estimate the margin in MNIST, CIFAR10 datasets via user studies. Their results suggest that for L_∞ perturbations larger than what is typically used in practice ($\epsilon = 0.1$), CIFAR10 doesn't not satisfy the margin condition. Together with our results, this shows that for such large perturbations, adversarial training will result in models with low standard accuracy. However, it is still unclear if the benchmark datasets satisfy the margin condition for ϵ typically used in practice. We believe answering this question can help us understand if it possible to obtain a truly robust model, without compromising on standard accuracy.

5.3 Standard training with increasing model complexity

Before we conclude the section, we show how our results from Theorem 3 can be used to explain an interesting phenomenon observed by Madry et al. [2017]: even with standard risk minimization, complex networks result in more robust classifiers than simple networks. Define the standard and adversarial training objectives as

$$\textbf{(standard)}\ \min_{f\in\mathcal{F}} R(f),$$

$$\textbf{(adversarial)}\ \min_{f\in\mathcal{F}} R(f) + \lambda G_{\text{adv}}(f).$$

Let $\mathcal{F}$ be a small function class, such as the set of functions which can be represented using a particular neural network architecture. As we increase the complexity of $\mathcal{F}$, we expect the minimizer of $R(f)$ to move closer to a Bayes optimal classifier. Assuming the margin condition is satisfied, from Theorem 3 we know that the minimizer of the adversarial training objective is also a Bayes optimal classifier. So as we increase the complexity of $\mathcal{F}$ we expect the joint risk $R(f) + \lambda G_{\text{adv}}(f)$ to go down. Conversely, a similar explanation can be used to explain the phenomenon that performing adversarial training on increasingly complex networks results in classifiers with better standard risk. Figures 3, 4 illustrate the two phenomena on MNIST and CIFAR10 datasets. To ensure the margin condition is at least approximately satisfied, we use small perturbations in these experiments. More details about the hyper-parameteres used in the experiments can be found in the Appendix.

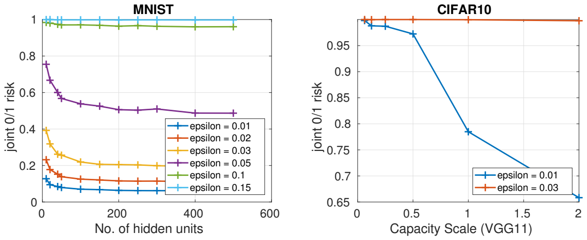


Figure 3: Behavior of joint 0/1 risk (*i.e.*, standard + adversarial risk) of models obtained through standard training, as we increase the model capacity.

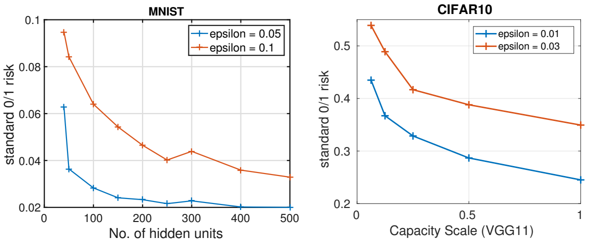


Figure 4: Behavior of standard 0/1 risk of models obtained through adversarial training (with $\lambda = 1$), as we increase the model capacity.

We conclude the discussion by pointing out that in practice we optimize empirical risks instead of population risks. So our explanations above are accurate only for smaller hypothesis spaces, where empirical risks and the corresponding population risks have similar landscapes.

6 Importance of Adversarial Training

Recall, in Section 4 we studied the properties of adversarial training when the base classifier is Bayes Optimal. In particular, in Theorem 1 we showed that the minimizers of adversarial training objective (4) are Bayes optimal classifiers, which are also the minimizers of standard risk. This naturally leads us to the following question: Do we really need adversarial training? Will standard training suffice to learn robust classifiers? In this section we show that standard risk minimization alone doesn't guarantee robust classifiers.

We first consider the setting where there is a single Bayes decision rule. In this setting, Theorem 1 shows that when the hypothesis class $\mathcal{F}$ is the set of all measurable functions, there is no need for adversarial training. However, in practice, we never optimize over the space of all measurable functions due to the finite amount of data available to us. Instead, we choose a small hypothesis class (such as the set of linear separators) apriori. In Section 6.2 we show that standard risk minimization over restricted hypothesis classes can result in classifiers with low standard risk but high adversarial risk.

In Section 6.3, we consider the setting where there are multiple Bayes decision rules. For instance, when the data is separable or lies in a low-dimensional manifold, Bayes decision rule is not unique. In this setting, even if one has access to unlimited data (which allows us to optimize over the space of all measurable functions), we show that there is a need for adversarial training. Although all the Bayes decision rules have the same standard risk, they can differ on adversarial risk. In such cases, it is impossible to distinguish these classifiers using standard risk. As a result, one needs to perform adversarial training to learn a robust Bayes decision rule.

We study these questions theoretically using a mixture model where the data for each class is generated from a different mixture component. The distribution of $\mathbf{x}$ conditioned on y follows a normal distribution: $\mathbf{x}|y \sim \mathcal{N}(y\mathbf{w}^*, \sigma^2\mathcal{I}_d)$, where $\mathcal{I}_d \in \mathbb{R}^{d\times d}$ is the identity matrix and $P(y=1) = P(y=-1) = \frac{1}{2}$. Note that in this setting $\mathbf{x} \mapsto \mathbf{x}^T\mathbf{w}^*$ is a Bayes optimal classifier. Morever there is a unique Bayes decision rule.

6.1 Calibration of Standard and Adversarial Risk

Firstly, we explore if the two risks are *calibrated*, *i.e. does approximately minimizing the standard risk always lead to small adversarial risk?* Suppose the mean of the Gaussian components is k-sparse; that is, $\mathbf{w}^*$ has k non-zero entries. Then the Bayes optimal classifier $\mathbf{x} \mapsto \mathbf{x}^T\mathbf{w}^*$ depends only on a few features and there are a lot of irrelevant features. The following result shows that there exist linear separators which achieve near-optimal classification accuracy, but have a high adversarial risk, even for a L_∞ adversarial perturbation of size $\frac{1}{\sqrt{d-k}}$.

Theorem 5. *Let $\mathbf{w}^*$ be k-sparse with non-zeros in the first k coordinates. Let $\mathbf{w} \in \mathbb{R}^d$ be a linear separator such that $\mathbf{w}_{1:k} = \mathbf{w^*}_{1:k}$, $\mathbf{w}_{k+1:d} = [\frac{\pm 1}{\sqrt{d-k}}, \dots, \frac{\pm 1}{\sqrt{d-k}}]$. Then, there exists a constant C such that if $\|\mathbf{w}^*\|_2 \geq C$ and $\sigma = 1$, the excess risk of $f_\mathbf{w}(\mathbf{x}) = \mathbf{x}^T\mathbf{w}$ is small; that is, $R_{0-1}(f_\mathbf{w}) - R^*_{0-1} \leq 0.02$, where R^*_{0-1} is the risk of the Bayes optimal classifier. However, even for a small enough perturbation $\epsilon \geq \frac{2\|\mathbf{w}^*\|_2^2}{\sqrt{d-k}}$ w.r.t L_∞ norm, the adversarial risk satisfies*

$$R_{adv,0-1}(f_\mathbf{w}) \geq 0.95,$$

where the base classifier $g(\mathbf{x})$ is equal to $sign(\mathbf{x}^T\mathbf{w}^*)$.

Note that the constructed classifier $\mathbf{w}$ has very small weights on irrelevant features. Hence the classification error is low *but not minimal.* But since there are a lot of such irrelevant features, there exist adversarial perturbations which don't change the prediction of Bayes classifer, but change the prediction of $\mathbf{w}$.

6.2 Optimizing Standard Risk over Restricted Function Class

Next, we study the effect of minimizing the standard risk over restricted function classes. Consider the restricted hypothesis class of all vectors which are nonzero in the top-k co-ordinates: $W_k = \{\mathbf{w} \in \mathbb{R}^d | \mathbf{w}(i) = 0 \ \forall i > k\}$. Our next result shows that the *exact* minimizer of standard risk over this restricted hypothesis class need *not* be the minimizer of the adversarial risk over this class, even for perturbations as small as $\frac{1}{\sqrt{d}}$.

Theorem 6. *Consider the gaussian mixture model with $\mathbf{w}^* = [\frac{1}{\sqrt{d/2}}, \frac{1}{\sqrt{d/2}}, \dots, \frac{1}{\sqrt{d/2}}]^T$, $\sigma = 1$ and let $\tilde{\mathbf{w}} = \text{argmin}_{\mathbf{w}\in W_{d/2}} R_{0-1}(f_\mathbf{w})$ be the minimizer of the standard risk when restricted to $W_{d/2}$. Then, even for a small enough perturbation of $\epsilon \geq \frac{C}{\sqrt{d}}$ w.r.t. L_∞ norm, we have that*

$$R_{0-1}(f_{\tilde{\mathbf{w}}}) - R_{0-1}(f_{\mathbf{w}^*}) < 0.1 \quad but \quad R_{adv,0-1}(f_{\tilde{\mathbf{w}}}) > 0.95,$$

where $R_{adv,0-1}(\cdot)$ is measured w.r.t. $g(\mathbf{x}) = sign(\mathbf{x}^T\mathbf{w}^*)$.

6.3 Multiple Bayes Decision Rules

In this section, we consider the setting where there could be multiple Bayes optimal decision rules. We consider the question of whether different Bayes optimal solutions have different adversarial risks, and whether standard risk minimization gives us robust Bayes optimal solutions.

Suppose our data comes from low dimensional Gaussians embedded in a high-dimensional space, *i.e.* suppose $\|\mathbf{w}^*\|_0 = k \ll d$ and the covariance matrix D of the conditional distributions $\mathbf{x}|y$ is diagonal with i^{th} diagonal entry $D_{ii} = \sigma^2$ if $\mathbf{w}^*_i \neq 0, 0$ otherwise. Notice that in this model any classifier $\tilde{\mathbf{w}}$ such that $\tilde{\mathbf{w}}_{1:k} = \mathbf{w^*}_{1:k}$ is a Bayes optimal classifier. Observe the subtle difference between this setting and sparse linear model. In particular, in the previous example, the data is inherently high-dimensional, but with only a few relevant discriminatory features; on the contrary, here the data lies on a low dimensional manifold of a high dimensional subspace.

In this setting, we study the adversarial risk of classifiers obtained through minimization of $R(f_\mathbf{w})$ using iterative methods such as gradient descent.

Theorem 7. *Let $\mathbf{w}^*$ be such that $\|\mathbf{w}^*\|_2 \geq C$, for some constant C. Let $\epsilon \geq \frac{2\|\mathbf{w}^*\|_2^2}{\sqrt{d-k}}$ and ℓ be any convex calibrated surrogate loss $\ell(f_\mathbf{w}(\mathbf{x}), y) = \phi(y\mathbf{w}^T\mathbf{x})$. Then gradient descent on $R(f_\mathbf{w})$ with random initialization using a Gaussian distribution with covariance $\frac{1}{\sqrt{d-k}}\mathcal{I}_d$ converges to a point $\hat{\mathbf{w}}_{GD}$ such that with high* *probability,*

$$R_{0,1}(f_{\hat{\mathbf{w}}_{GD}}) = 0 \quad \textit{but} \quad R_{adv,0-1}(f_{\hat{\mathbf{w}}_{GD}}) \geq 0.95,$$

where $R_{adv,0-1}(\cdot)$ *is the adversarial risk measured w.r.t.* $\mathbf{w}^*$*.*

Note that Theorem 7 raises the vulnerability of standard risk minimization by showing that it can lead to Bayes optimal solutions which have high adversarial risk. Moreover, observe that increasing d results in classifiers that are less robust; even a $O(1/\sqrt{d-k})$ perturbation can create adversarial examples with respect to $\mathbf{w}^*$. All our results in this section show that standard risk minimization is inherently insufficient in providing robustness. This suggests the need for adversarial training.

7 Regularization properties of Adversarial Training

In this section, we study the regularization properties of the adversarial training objective in Equation (4). Specifically, we show that the adversarial risk $R_{\text{adv}}(f)$, effectively acts as a regularizer which biases the solution towards certain classifiers. The following Theorem explicitly shows this regularization effect of adversarial risk.

Theorem 8. *Let* $\|.\|_*$ *be the dual norm of* $\|.\|$*, which is defined as:* $\|\mathbf{z}\|_* = \sup_{\|\mathbf{x}\|=1} \mathbf{z}^T\mathbf{x}$*. Suppose* ℓ *is the logistic loss and suppose the classifier* f *is differentiable a.e. Then for any* $\epsilon \geq 0$ *the adversarial training objective (4) can be upper bounded as*

$$R(f) + \lambda R_{adv}(f) \leq$$
$$R(f) + \lambda \min\left\{\epsilon\mathbb{E}\left[\sup_{\|\boldsymbol{\delta}\|\leq\epsilon} \|\nabla f(\mathbf{x}+\boldsymbol{\delta})\|_*\right], 2\|f-g\|_\infty\right\},$$

where $\|f-g\|_\infty = \sup_{\mathbf{x}} |f(\mathbf{x}) - g(\mathbf{x})|$*.*

Although the above Theorem only provides an upper bound, it still provides insights into the regularization effects of adversarial risk. It shows that adversarial risk effectively acts as a regularization term biasing the optimization towards two kinds of classifiers: 1) classifiers that are smooth with small gradients and 2) classifiers that are pointwise close to the base classifier $g(\mathbf{x})$. We now compare the regularization effect of adversarial risk in objective (4) with the regularization effect of existing notion of adversarial risk.

Theorem 9. *Suppose* ℓ *is the logistic loss and suppose the classifier* f *is differentiable a.e. Then for any* $\epsilon \geq 0$ *the adversarial training objective (7) can be upper bounded as*

$$R(f) + \lambda G_{adv}(f) \leq R(f) + \lambda\epsilon\mathbb{E}\left[\sup_{\|\boldsymbol{\delta}\|\leq\epsilon} \|\nabla f(\mathbf{x}+\boldsymbol{\delta})\|_*\right].$$

Moreover, for linear classifiers $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$*, the adversarial training objective (7) can be upper and lower bounded as*

$$R(f_{\mathbf{w}}) + \lambda G_{adv}(f_{\mathbf{w}}) \leq R(f_{\mathbf{w}}) + \lambda\epsilon\|\mathbf{w}\|_*, \tag{8}$$

$$R(f_{\mathbf{w}}) + \lambda G_{adv}(f_{\mathbf{w}}) \geq R(f_{\mathbf{w}}) + \left(\frac{\lambda\epsilon}{2}R_{0-1}(f_{\mathbf{w}})\right)\|\mathbf{w}\|_*.$$

Comparing Theorems 8, 9, we can see that the major difference between the two adversarial risks is that the existing definition doesn't necessarily bias the optimization towards the base classifier $g(\mathbf{x})$, whereas the new definition certainly biases the optimization towards $g(\mathbf{x})$.

For linear classifiers, the above Theorem provides a tight upper bound and shows that adversarial training using objective (7) essentially acts as a regularizer which penalizes the dual norm of $\mathbf{w}$. In a related work, Xu et al. [2009] focus on linear classifiers with hinge loss, and show that under separability conditions on the data and certain additional constraint on perturbations, the robust objective is equivalent to the regularized objective.

8 Summary and Future Work

In this work, we identified the inaccuracies with the existing definition of adversarial risk and proposed a new definition of adversarial risk which fixes these inaccuracies. We analyzed the properties of minimizers of the resulting adversarial training objective and showed that Bayes optimal classifiers are its minimizers and that there is no trade-off between adversarial and standard risks. We also study the existing definition of adversarial risk, its relation to the new definition, and identify conditions under which its minimizers are Bayes optimal. Our analysis highlights the importance of margin for Bayes optimality of its minimizers.

An important direction for future work would be to design algorithms for minimization of the new adversarial training objective. One can consider two different approaches in this direction: 1) assuming we have black box access to the base classifier, one could design efficient optimization techniques which make use of the black box. 2) assuming we have access to an approximate base classifier (*e.g.,* some complex model which is pre-trained on a lot of labeled data or a "teacher" network), one could use this classifier as a surrogate for the base classifier, to optimize the adversarial training objective.

9 Acknowledgements

We acknowledge the support of NSF via IIS-1149803, IIS-1664720, DARPA via FA87501720152, and PNC via the PNC Center for Financial Services Innovation.

References

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *arXiv preprint arXiv:1608.04644*, 2016.

Andrew Ilyas, Ajil Jalal, Eirini Asteri, Constantinos Daskalakis, and Alexandros G Dimakis. The robust manifold defense: Adversarial training using generative models. *arXiv preprint arXiv:1712.09196*, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Anish Athalye and Ilya Sutskever. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

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

A. Athalye, N. Carlini, and D. Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. *ArXiv e-prints*, February 2018.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *arXiv preprint arXiv:1802.04034*, 2018.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

J. Zico Kolter and Eric Wong. Provable defenses against adversarial examples via the convex outer adversarial polytope. *CoRR*, abs/1711.00851, 2017. URL http://arxiv.org/abs/1711.00851.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Mądry. Adversarially robust generalization requires more data. *arXiv preprint arXiv:1804.11285*, 2018.

Sébastien Bubeck, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*, 2018.

Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508, 2018.

A. Fawzi, H. Fawzi, and O. Fawzi. Adversarial vulnerability for any classifier. *ArXiv e-prints*, February 2018.

Jean-Yves Franceschi, Alhussein Fawzi, and Omar Fawzi. Robustness of classifiers to uniform l and gaussian noise. *arXiv preprint arXiv:1802.07971*, 2018.

Mahmood Sharif, Lujo Bauer, and Michael K Reiter. On the suitability of lp-norms for creating and preventing adversarial examples. *arXiv preprint arXiv:1802.09653*, 2018.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. There is no free lunch in adversarial robustness (but there are unexpected benefits). *arXiv preprint arXiv:1805.12152*, 2018.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10 (Jul):1485–1510, 2009.