Fundamental Tradeoffs in Distributionally Adversarial Training

Mohammad Mehrabi [*1] **Adel Javanmard** [*1] **Ryan A. Rossi** [2] **Anup B. Rao** [2] **Tung Mai** [2]

Abstract

Adversarial training is among the most effective techniques to improve robustness of models against adversarial perturbations. However, the full effect of this approach on models is not well understood. For example, while adversarial training can reduce the adversarial risk (prediction error against an adversary), it sometimes increase standard risk (generalization error when there is no adversary). In this paper, we focus on *distribution perturbing* adversary framework wherein the adversary can change the test distribution within a neighborhood of the training data distribution. The neighborhood is defined via Wasserstein distance between distributions and the radius of the neighborhood is a measure of adversary's manipulative power. We study the tradeoff between standard risk and adversarial risk and derive the Pareto-optimal tradeoff, achievable over specific classes of models, in the infinite data limit with features dimension kept fixed. We consider three learning settings: 1) Regression with the class of linear models; 2) Binary classification under the Gaussian mixtures data model, with the class of linear classifiers; 3) Regression with the class of random features model (which can be equivalently represented as two-layer neural network with random first-layer weights). We show that a tradeoff between standard and adversarial risk is manifested in all three settings. We further characterize the Pareto-optimal tradeoff curves and discuss how a variety of factors, such as features correlation, adversary's power or the width of two-layer neural network would affect this tradeoff.

1. Introduction

Modern machine learning algorithms, and in particular deep neural networks, have demonstrated breakthrough empirical performance, and have been deployed in a multitude of applications domains ranging from visual object classification to speech recognition, robotics, natural language processing and healthcare. The common practice to train these models is by empirical loss minimization on the training data. Nonetheless, it has been observed that the resulting models are surprisingly vulnerable to minute discrepancies between the test and the training data distributions. There are several well documented examples of such behavior in computer vision and image processing where small imperceptible manipulations of images can significantly compromise the performance of the state-of-the-art classifiers (Szegedy et al., 2014; Biggio et al., 2013). Other examples include well-designed malicious content like malware which can be labeled legitimate by the classifier and harm the system (Chen et al., 2017; Papernot et al., 2017), or adversarial attacks on speech recognition systems, such as GoogleNow or Siri, which consists in voice commands that are incomprehensible or even completely inaudible to human and can still control the systems (Carlini et al., 2016; Vaidya et al., 2015; Zhang et al., 2017). It is evident that in practice such vulnerability can have catastrophic consequences.

By studying adversarial samples, one can in turn improve the robustness of machine learning algorithms against adversarial attacks. In the past few years, there has been a significant research on generating various adversarial samples (Carlini & Wagner, 2017; Athalye et al., 2018; Goodfellow et al., 2015b; Papernot et al., 2016a) and defenses (Madry et al., 2018a; Cisse et al., 2017; Papernot et al., 2016b). Among the considerable effort to improve the adversarial robustness of algorithms, adversarial training is one of the most effective techniques. Adversarial training is often formulated as a minimax optimization problem, where the inner maximization aims to find an adversarial example that maximizes a predictive loss function, while the outer minimization aims to train a robust estimator using the generated adversarial examples (Goodfellow et al., 2015a; Kurakin et al., 2016; Madry et al., 2018a; Raghunathan et al., 2018; Wong & Kolter, 2018).

While adversarial training techniques have been successful in improving the adversarial robustness of the models, their full effect on machine learning systems is not well understood. In particular, some studies (Madry et al., 2018a) observed that the robustness virtue of adversarial training

[*]Equal contribution [1]Department of Data Sciences and Operations, USC Marshall School of Business, University of Southern California, USA; [2]Adobe Research, USA. Correspondence to: Adel Javanmard <ajavanma@usc.edu>.

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

comes at the cost of worsening the performance on natural unperturbed inputs, i.e, increasing generalization error. However, (Tsipras et al., 2018) observes empirically that when there are very few training data, adversarial training can help with reducing the generalization error. Complicating matters further, (Raghunathan et al., 2019) argues that additional unlabeled data can mitigate the tension between adversarial risk (predictive performance against adversarial perturbations) and the standard risk (predictive performance when there is no adversary, a.k.a generalization error). These observations raise the following important question regarding adversarial training:

Is there a 'fundamental' tradeoff between adversarial risk and standard risk? Or do there exist models that are optimal with respect to both of these measures? What are the roles of different factors, such as adversary's power, problem dimension and the complexity of the model class (e.g., number of neurons) in the interplay between standard risk and adversarial risk?

Here, by 'fundamental tradeoff' we mean a tradeoff that holds given unlimited computational power and infinite training data to train a model. In this work, we answer these questions for adversarial distribution shifts, where the adversary can shift the test data distribution, making it different from the training data distribution. The test data distribution can be an arbitrary but fixed distribution in a neighborhood of the training data distribution and the radius of this neighborhood is a measure of adversary's power.

Contributions. We next summarize our contributions in this paper:

- We characterize the fundamental tradeoff between standard risk and adversarial risk for distributionally adversarial training for the settings of linear regression and binary classification (under a Gaussian mixtures model). We focus on infinite data limit $(n \to \infty)$ with finite feature dimension (d) and hence our analysis is at population level. The fundamental tradeoff is characterized by studying the Pareto optimal fronts for the achievability region in the two dimensional standard risk-adversarial risk region. The Pareto optimal front consists in the set of estimators for which one cannot decrease both standard and adversarial risk by deviating from these estimators. Similar tradeoffs have been derived for linear regression setting with norm bounded adversarial perturbation and isotropic Gaussian features (Javanmard et al., 2020). Here we focus on distribution perturbing adversaries and consider general anisotropic Gaussian features.
- For the binary classification we consider a Gaussian mixtures model with general feature covariance and a distribution perturbing adversary, where the perturbation is measured in terms of the Wasserstein metric with general ℓ_r norm. (We refer to Sections 2.2 and 3.2 for further details and formal definitions). Our analysis shows how the fundamental tradeoff between standard and adversarial risk is impacted by a variety of factors, such as adversary's power, feature dimension, features correlation and the choice of ℓ_r perturbation norm. An interesting observation is that for $r = 2$ the tradeoff between standard and adversarial risk vanishes. In other words, there exists a model which achieve both the optimal standard risk and the optimal adversarial risk.
- We also study the Pareto optimal tradeoffs between the standard and adversarial risks for the problem of learning an *unknown function* over the d-dimensional sphere using random features model. This can be represented as linear models with N random nonlinear features of the form $\sigma(w_a^{\mathsf{T}} x)$, $1 \le a \le N$, with $\sigma(\cdot)$ a nonlinear activation. Equivalently this can be characterized as fitting a two-layer neural network with random first-layer. Building upon approximation formula for adversarial risk, we study the effect of network width N on the tradeoff between standard and adversarial risks.

1.1. Further related work

Very recent work (Javanmard & Soltanolkotabi, 2020; Taheri et al., 2020) have focused on binary classification, under Gaussian mixtures model and proposed a precise characterization of the standard and adversarial risk achieved by a specific class of adversarial training approach (Tsipras et al., 2018; Madry et al., 2018b). These work consider an asymptotic regime where the sample size grows in proportion to the problem dimension d and focus on norm bounded adversarial perturbation. In comparison, we consider a fixed d, infinite n setting and consider distribution perturbing adversary. Also we focus on fundamental tradeoffs achieved by any linear classifier, while (Javanmard & Soltanolkotabi, 2020; Taheri et al., 2020) work with a specific class of saddle point estimator. The other work (Dobriban et al., 2020) also considers norm bounded adversarial perturbation for the classification problem and studies the optimal ℓ_2 and ℓ_∞ robust linear classifiers assuming access to the class averages. Furthermore, it also studies the tradeoff between standard and robust accuracies from a Bayesian perspective by contrasting this optimal robust classifier with the Bayes optimal classifier in a non-adversarial setting.

2. Problem formulation

In a classic supervised learning setting, a learner is given n pairs of data points $\{z_i := (x_i, y_i)\}_{i=1:n}$ with $x_i \in \mathbb{R}^d$ representing features vectors and y_i the response variables (or labels). The common assumption in supervised machine learning is that the data points z_i are drawn independently and identically from some probability measure $\mathbb{P}_Z$ defined

over the space $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. Given this training data, the learner would like to fit a parametric function f_θ with $\theta \in \mathbb{R}^d$ to predict the response (label) on new points x.

A common practice to model fitting is through the empirical risk minimization:

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{j=1}^{n} \ell(\theta; (x_j, y_j)) \,, \tag{1}$$

with $\ell(\theta; (x, y)) := \widetilde{\ell}(f_\theta(x), y)$ and $\widetilde{\ell}$ being a loss function which captures the discrepancy between the estimated value $f_\theta(x)$ and the true response value y. The performance of the model is then measured in terms of *standard risk* (a.k.a. generalization error), defined as

$$\mathsf{SR}(\theta) := \mathbb{E}_{z=(x,y)\sim\mathbb{P}_Z} \left[\ell(\theta; (x, y))\right] \,. \tag{2}$$

Standard risk is a population risk and quantifies the expected error on new data points drawn from the same distribution as the training data.

Although the empirical risk minimization is a widely used approach for model learning, it is well known that the resulting models can be highly vulnerable to adversarial perturbations of their inputs, known as adversarial attacks. We next discuss the adversarial setting and two common adversary models that are considered in literature.

2.1. Adversarial setting

The adversarial setting can be perceived as a game between the learner and the adversary. Given access to the training data, drawn i.i.d from a common distribution $\mathbb{P}_Z$, the learner chooses a model θ. Depending on the adversary's budget ε, the adversary chooses a test data point $(\tilde{x}, \tilde{y})$ that can deviate from a typical test point according to one of the following models. The performance of model θ is then measured in terms of predicting $\tilde{y}$ given the perturbed input $\tilde{x}$.

Norm-bounded perturbations. In this setting, $\tilde{y} = y$ (no perturbation on the response) and $\tilde{x} = x + \delta$ where δ can be an arbitrary vector from ℓ_r-ball of radius ε. The *adversarial risk* in this case is defined as

$$\mathsf{AR}(\theta) := \mathbb{E}_{(x,y)\sim\mathbb{P}_Z} \left[\sup_{\|\delta\|_{\ell_r} \le \varepsilon} \ell(\theta; (x + \delta, y)) \right] \,. \tag{3}$$

Distribution shift. In this setting, the adversary can shift the distribution of test data, making it different than the training distribution $\mathbb{P}_Z$. Specifically, $(\tilde{x}, \tilde{y}) \sim \mathbb{Q}$ where $\mathbb{Q} \in \mathcal{U}_\varepsilon(\mathbb{P}_Z)$ denotes an ε- neighborhood of the distribution $\mathbb{P}_Z$. A popular choice of this neighborhood is via the Wasserstein distance, which is formally defined below. In this case, the *adversarial risk* is defined as

$$\mathsf{AR}(\theta) := \sup_{\mathbb{Q}\in\mathcal{U}_\varepsilon(\mathbb{P}_Z)} \mathbb{E}_{(\tilde{x},\tilde{y})\sim\mathbb{Q}} \left[\ell(\theta; (\tilde{x}, \tilde{y}))\right] \,. \tag{4}$$

Note that this is a strong notion of adversary as the perturbation is chosen *after* observing both the model θ and data point (x, y) (in norm-bounded perturbation model) or the training data distribution $\mathbb{P}_Z$ (in the distribution shift model). The distribution perturbing adversary is a common model in a multitude of application domains and has already been adopted by several work including (Staib & Jegelka, 2017; Dong et al., 2020; Pydi & Jog, 2020).

Our primary focus on this work is on the distribution shift adversary model with Wasserstein metric to measure the distance between distributions. The next section provides a brief background on the Wasserstein robust loss which will be used later in our work.

2.2. Background on Wasserstein robust loss

Let $\mathcal{Z}$ be a metric space endowed with metric $d : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\ge 0}$. Denote by $\mathcal{P}(\mathcal{Z})$ the set of all Borel probability measures on $\mathcal{Z}$. For a $\mathbb{Q}$-measurable function f, the $\mathcal{L}^p(\mathbb{Q})$-norm of f is defined as

$$\|f\|_{\mathbb{Q},p} := \begin{cases} \left(\int_{\mathcal{Z}} |f|^p \, \mathrm{d}\mathbb{Q}\right)^{1/p} & \text{for } p \in [1, \infty) \\ \mathbb{Q}\text{-ess sup}_{z\in\mathcal{Z}} \; |f(z)| & \text{for } p = \infty \end{cases} \tag{5}$$

For two distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{Z})$ the Wasserstein distance of order p is given by

$$W_p(\mathbb{P}, \mathbb{Q}) := \inf_{\pi \in \mathrm{Cpl}(\mathbb{P},\mathbb{Q})} \|d\|_{\pi,p} \,, \tag{6}$$

where the coupling set $\mathrm{Cpl}(\mathbb{P}, \mathbb{Q})$ denotes the set of all probability measures π on $\mathcal{Z} \times \mathcal{Z}$ with the first marginal $\pi_1 := \pi(\cdot \times \mathcal{Z}) = \mathbb{P}$ and the second marginal $\pi_2 := \pi(\mathcal{Z} \times \cdot) = \mathbb{Q}$.

We use the Wasserstein distance to define the neighborhood set $\mathcal{U}_\varepsilon$ in the distribution shift adversary model. Namely,

$$\mathcal{U}_\varepsilon(\mathbb{P}_Z) := \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z}) : \; W_p(\mathbb{P}_Z, \mathbb{Q}) \le \varepsilon\} \,. \tag{7}$$

In this case we refer to $\mathsf{AR}(\theta)$ given by (4) as *Wasserstein adversarial risk*. Note that this notion involves a maximization over distributions $\mathbb{Q} \in \mathcal{P}(\mathcal{Z})$ which can be daunting. However, an important result from distributional robust optimization which we also use in our characterization of $\mathsf{AR}(\theta)$ is that the strong duality holds for this problem under general conditions. The dual problem of (4) is given by

$$\begin{cases} \min_{\gamma \ge 0} \left\{\gamma \varepsilon^p + \mathbb{E}_{\mathbb{P}_Z}[\phi_\gamma(\theta; z)]\right\} \,, & p \in [1, \infty), \\ \mathbb{E}_{z\sim\mathbb{P}_Z}\left[\sup_{\tilde{z}\in\mathcal{Z}} \{\ell(\theta; \tilde{z}) : \; d(z, \tilde{z}) \le \varepsilon\}\right] & p = \infty \,. \end{cases} \tag{8}$$

Here $\phi_\gamma(\theta; z)$ is the robust surrogate for the loss function $\ell(\theta; z)$ and is defined as

$$\phi_\gamma(\theta; z_0) := \sup_{z\in\mathcal{Z}} \{\ell(\theta; z) - \gamma d^p(z, z_0)\} \,. \tag{9}$$

For $p \in [1,\infty)$ it is shown that strong duality holds if either $\mathbb{P}_Z$ has finite support or $\mathcal{Z}$ is a Polish space (Gao & Kleywegt, 2016). For $p = \infty$, Lemma EC.2 in (Gao et al., 2020) shows that strong duality holds if $\mathbb{P}_Z$ has finite support.

Remark 2.1. *It is worth noting that the Wasserstein adversary model is stronger than and generalizes the norm-bounded perturbation model. In particular, for any $p \in [1,\infty]$,*

$$\mathbb{E}_{(x,y)\sim\mathbb{P}_Z}\left[\sup_{\|\delta\|_{\ell_r}\le\varepsilon} \ell(\theta;(x+\delta,y))\right] \le \sup_{\mathbb{Q}\in\mathcal{U}_\varepsilon(\mathbb{P}_Z)} \mathbb{E}_{(\tilde{x},\tilde{y})\sim\mathbb{Q}}\left[\ell(\theta;(\tilde{x},\tilde{y}))\right],$$

where $\mathcal{U}_\varepsilon(\mathbb{P}_Z)$ is given by (7) with the Wasserstein distance defined with respect to the ℓ_r distance which is also used in the definition of norm-bounded perturbation model. Equality holds for $p = \infty$. We refer to (Staib & Jegelka, 2017, Proposition 3.1) and (Pydi & Jog, 2020, Corollary 2.1, Theorem 1) for the proof and further explanation.

2.2.1. Regularization effect of Wasserstein adversarial loss

It is clear from the definition that $\mathsf{AR}(\theta) \ge \mathsf{SR}(\theta)$ for any model θ. Understanding the tradeoff between standard and adversarial risks is intimately related to the gap $\mathsf{AR}(\theta) - \mathsf{SR}(\theta)$. The gap between the Wasserstein adversarial loss and the standard loss has been studied in several settings in the context of distributionally robust optimization (DRO) (Bartl et al., 2020; Gao et al., 2020). In particular, (Bartl et al., 2020; Gao et al., 2020) introduced the notion of *variation of the loss*, denoted as $\mathcal{V}(\ell)$, as a measure of the magnitude change in the expected loss when the data distribution is perturbed, and showed that the Wasserstein adversarial loss is closely related to regularizing the nominal loss by the variation $\mathcal{V}(\ell)$ regularizer. The formal definition of the variation of loss, recalled from (Gao et al., 2020), is given below.

Definition 2.2. *(Variation of the loss). Suppose that $\mathcal{Z}$ is a normed space with norm $\|\cdot\|$. Let ℓ be a continuous function on $\mathcal{Z}$. Also assume that $\nabla_z \ell$ exists $\mathbb{P}$-almost everywhere. The variation of loss ℓ with respect to $\mathbb{P}$ is defined as*

$$\mathcal{V}_{\mathbb{P},q}(\ell) := \begin{cases} \| \|\nabla_z \ell\|_* \|_{\mathbb{P},q} & q \in [1,\infty), \\ \mathbb{P}_Z\text{-esssup}_{z\in\mathcal{Z}} \sup_{\tilde{z}\neq z} \frac{(\ell(\tilde{z})-\ell(z))_+}{\|\tilde{z}-z\|}, & q = \infty. \end{cases} \tag{10}$$

Here $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$ and we recall that $\|\cdot\|_{\mathbb{P},q}$ is the $\mathcal{L}^q(\mathbb{P})$-norm given by (5).

The following proposition from (Bartl et al., 2020; Gao et al., 2020) states that the variation of loss captures the first order term of the gap between Wasserstein adversarial risk and standard risk for small ε.

Proposition 2.3. *Suppose that the loss $\ell(\theta;z)$ is differentiable in the interior of $\mathcal{Z}$ for every θ, and $\nabla_z \ell$ is continuous on $\mathcal{Z}$. When $p \in (1,\infty)$, assume that there exists $M, L \ge 0$ such that for every θ and $z, \tilde{z} \in \mathcal{Z}$,*

$$\|\nabla_z \ell(\theta;\tilde{z}) - \nabla_z \ell(\theta;z)\|_* \le M + L\|\tilde{z}-z\|^{p-1}.$$

When $p = \infty$, assume instead that there exists $M \ge 0$ and $\delta_0 > 0$ such that for every θ and $z, \tilde{z} \in \mathcal{Z}$ with $\|\tilde{z}-z\| < \delta_0$, we have

$$\|\nabla_z \ell(\theta;\tilde{z}) - \nabla_z \ell(\theta;z)\|_* \le M.$$

Then, there exists $\bar{\varepsilon}$ such that for all $0 \le \varepsilon < \bar{\varepsilon}$ and all θ

$$\mathsf{AR}(\theta) - \mathsf{SR}(\theta) = \varepsilon \mathcal{V}_{\mathbb{P}_Z,q}(\ell) + O(\varepsilon^2), \tag{11}$$

where $\frac{1}{p} + \frac{1}{q} = 1$ and p is the order of Wasserstein distance in defining set $\mathcal{U}_\varepsilon(\mathbb{P}_Z)$ in the adversarial risk (4).

By virtue of Proposition 2.3, the Wasserstein adversarial risk can be perceived as a regularized form of the standard risk with regularization given by the variation of the loss. Nonetheless, note that this is only an approximation which captures the first order terms for small adversary's power ε. (See also Remark 8 in (Bartl et al., 2020) for an upper bound on the gap, up to second order terms in ε.)

In this paper, we consider the settings of linear regression and binary classification. For these settings, only for the special case of $p = 1$ (1-Wasserstein) and when the loss is Lipschitz and its derivative converges at ∞, it is shown that the gap (11) is linear in ε and therefore is precisely characterized as $\varepsilon \mathcal{V}_{\mathbb{P}_Z,q}(\ell)$. However, as we will consider more common losses for these settings, namely quadratic loss for linear regression and 0-1 loss for classification, such characterization does not apply to our settings and requires a direct derivation of adversarial risk. Later, in Section 3.3 we use the result of Proposition 2.3 to study the tradeoff between SR and AR in the problem of learning an unknown function over the d-dimensional sphere $\mathbb{S}^{d-1}$.

3. Main results

In this paper we focus on the distribution perturbing adversary and aim at understanding the fundamental tradeoff between standard risk and adversarial risk, which holds regardless of computational power or the size of training data. We consider 2-Wasserstein distance ($p = 2$) with the metric $d(z,\tilde{z})$ defined as

$$d(z,\tilde{z}) = \|x - \tilde{x}\|_{\ell_2} + \infty \cdot \mathbb{I}\{y \neq \tilde{y}\}, \tag{12}$$

for $z = (x,y)$ and $\tilde{z} = (\tilde{x},\tilde{y})$. Therefore, the adversary with a finite power ε can only perturb the distribution of the

input feature x, but not y. Otherwise, the distance $d(z, \tilde{z})$ becomes infinite and so the Wasserstein distance between the the data distribution $\mathbb{P}_Z$ and the adversary distribution $\mathbb{Q}$, given by (6), also becomes infinite. It is worth noting that this choice of d is only for simplicity of presentations and our results in this section can be derived in a straightforward manner for distances that also allow perturbations on the y component.

The following remark relates the two types of adversary discussed in Section 2.1 and follows readily from the definition (6) and Equation (5).

Remark 3.1. *For distance $d(\cdot, \cdot)$ given by (12), the adversary model with norm bounded perturbations correspond to the distribution shifting adversary model with $p = \infty$ Wasserstein distance.*

3.1. Linear regression

We consider the class of linear models to fit to data with quadratic loss $\ell(z; \theta) = (y - x^\mathsf{T}\theta)^2$. Our first result is a closed form representation of the Wasserstein adversarial risk (4) in this case.

Proposition 3.2. *Consider the quadratic loss $\ell(z; \theta) = (y - x^\mathsf{T}\theta)^2$ and the distribution perturbing adversary with $\mathcal{U}_\varepsilon(\mathbb{P}_Z)$ given by (7) with $p = 2$ and the metric d given by (12). In this case the adversarial risk $\mathsf{AR}(\theta)$ admits the following form:*

$$\mathsf{AR}(\theta) = \left(\sqrt{\mathbb{E}_{\mathbb{P}_Z}[(y - x^\mathsf{T}\theta)^2]} + \varepsilon \|\theta\|_{\ell_2}\right)^2 . \tag{13}$$

To prove Proposition 3.2 we exploit the dual problem (8). We refer to Section A.1 for the proof of Proposition 3.2.

Pareto optimal curve. For the linear regression setting, note that the standard risk $\mathsf{SR}(\theta)$ and the adversarial risk $\mathsf{AR}(\theta)$ are convex functions of θ. (The latter is convex since $\mathbb{E}_\mathbb{Q}[(y - x^\mathsf{T}\theta)^2]$ is convex for any distribution $\mathbb{Q}$ and maximization preserves convexity.) Therefore, we can find (almost) all Pareto optimal points by minimizing a weighted combination of the two risk measures by varying the weight λ:

$$\theta_\lambda := \arg\min_\theta \lambda \mathsf{SR}(\theta) + \mathsf{AR}(\theta) \tag{14}$$

The Pareto optimal curve is then given by $\{(\mathsf{SR}(\theta_\lambda), \mathsf{AR}(\theta_\lambda)) : \lambda \geq 0\}$.

Theorem 3.3. *Consider the setting of Proposition 3.2 with $v := \mathbb{E}[yx]$, $\sigma_y^2 := \mathbb{E}[y^2]$, and $\Sigma := \mathbb{E}[xx^\mathsf{T}]$. Then the solution θ of optimization (14) is given either by (i) $\theta_\lambda = 0$ or (ii) $\theta_\lambda = (\Sigma + \gamma_* I)^{-1} v$, with γ_* the fixed point of the following two equations:*

$$\gamma = \frac{\varepsilon^2 + \varepsilon A}{1 + \lambda + \frac{\varepsilon}{A}} , \tag{15}$$

$$A = \frac{1}{\|(\Sigma + \gamma I)^{-1} v\|_{\ell_2}} \left(\sigma_y^2 + \left\|\Sigma^{1/2}(\Sigma + \gamma I)^{-1} v\right\|_{\ell_2}^2 - 2v^\mathsf{T}(\Sigma + \gamma I)^{-1} v\right)^{1/2} . \tag{16}$$

In case (i) we have $\mathsf{SR}(\theta_\lambda) = \mathsf{AR}(\theta_\lambda) = \sigma_y^2$. In case (ii) we have

$$\begin{aligned} \mathsf{SR}(\theta_\lambda) &= A_*^2 \left\|(\Sigma + \gamma_* I)^{-1} v\right\|_{\ell_2}^2 , \\ \mathsf{AR}(\theta_\lambda) &= (A_* + \varepsilon)^2 \left\|(\Sigma + \gamma_* I)^{-1} v\right\|_{\ell_2}^2 , \end{aligned} \tag{17}$$

where A_* is given by (16) when $\gamma = \gamma_*$.

The proof of Theorem 3.3 is given in Section A.2.

Corollary 3.4. *Suppose that data is generated according to linear model $y = x^\mathsf{T}\theta_0 + w$ with $w \sim \mathsf{N}(0, \sigma^2)$ and isotropic features satisfying $\mathbb{E}[xx^\mathsf{T}] = I_d$. Then the solution θ_λ of optimization (14) is given either by (i) $\theta_\lambda = 0$ or (ii) $\theta_\lambda = (1 + \gamma_*)^{-1}\theta_0$, where γ_* is the fixed point of the following two equations:*

$$\gamma = \frac{\varepsilon^2 + \varepsilon A}{1 + \lambda + \frac{\varepsilon}{A}} , \tag{18}$$

$$A = \left(\gamma^2 + (1 + \gamma)^2 \frac{\sigma^2}{\|\theta_0\|_{\ell_2}^2}\right)^{1/2} . \tag{19}$$

In case (i) we have $\mathsf{SR}(\theta_\lambda) = \mathsf{AR}(\theta_\lambda) = \sigma^2 + \|\theta_0\|_{\ell_2}^2$. In case (ii) we have

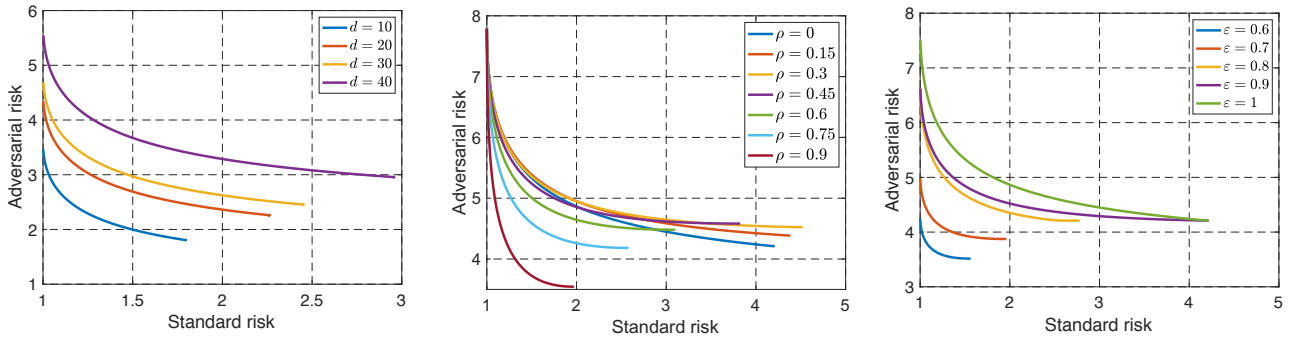
$$\mathsf{SR}(\theta_\lambda) = A_*^2 (1 + \gamma_*)^{-2} \|\theta_0\|_{\ell_2}^2 , \tag{20}$$

$$\mathsf{AR}(\theta_\lambda) = (A_* + \varepsilon)^2 (1 + \gamma_*)^{-2} \|\theta_0\|_{\ell_2}^2 , \tag{21}$$

where A_* is given by (19) when $\gamma = \gamma_*$.

The proof of Corollary 3.4 is provided in Section A.3.

Figure 1 shows the effect of various parameters on the Pareto optimal tradeoffs between adversarial (AR) and standard risks (SR) in linear regression setting. We consider data generated according to the linear model $y = x^\mathsf{T}\theta_0 + w$ with $w \sim \mathsf{N}(0, 1)$ and features x_i sampled i.i.d from $\mathsf{N}(0, \Sigma)$ where $\Sigma_{i,j} = \rho^{|i-j|}$. Figure 1a demonstrates the role of features dimension d on the Pareto optimal curve for the setting with $\rho = 0$ (identity covariance matrix), adversary's power $\varepsilon = 1$, and the entries of θ_0 generated independently from $\mathsf{N}(0, 1/40)$. Note that by Corollary 3.4, in the case of isotropic features, standard risk and adversarial risks depend



(a) Pareto optimal curve for several feature dimensions d with $\rho = 0$ and $\varepsilon = 1$.

(b) Pareto optimal curve for several feature dependency values ρ with $d = 10$, $\varepsilon = 1$.

(c) Pareto optimal curve for several adversary's power ε with $\rho = 0$ and $d = 10$.

Figure 1. The effect of feature dimension (d), dependency across features (ρ), and adversary's power (ε) on Pareto optimal tradeoff between adversarial (AR) and standard risks (SR) in linear regression setting.

on θ_0 only through its ℓ_2 norm. The variations in the Pareto-curve here is due to variations in $\|\theta_0\|_{\ell_2}$ as d changes.

Figure 1b investigates the role of dependency across features (ρ) in the optimal tradeoff between standard and adversarial risks. In this setting $d = 10$, $\varepsilon = 1$, and $\theta_0 \sim \frac{1}{\sqrt{d}}\mathsf{N}(0, I)$. As we see all the curves start from the same point. This can be easily verified by the result of Theorem 3.3: For the linear data model $y = x^\mathsf{T}\theta_0 + w$, we have $v = \Sigma\theta_0$ and at $\lambda = \infty$, the Pareto-optimal estimator is the minimizer of the standard risk, i.e. $\theta_{\lambda=\infty} = \theta_0$. Also by (15) we have $\gamma_* = 0$, and by (16) we obtain $A = \sigma / \|\theta_0\|_{\ell_2}$. Plugging these values in (17) we get $\mathsf{SR}(\theta_\infty) = \sigma^2$ and $\mathsf{AR}(\theta_\infty) = (\sigma + \varepsilon_0 \|\theta_0\|_{\ell_2})^2$. Therefore both metrics become independent of ρ at $\lambda = \infty$. Also looking at the right-most point of the Pareto-curves, corresponding to $\lambda = 0$, we see that as ρ increases from small to moderate values, this point moves upward-right, indicating that both standard and adversarial risks increase, but after some value of ρ, we start to see a reverse behavior, where standard and adversarial risks start to decrease.

Finally in Figure 1c we observe the effect of adversary's budget ε on the Pareto optimal curve. Here, $d = 10$, $\rho = 0$, and $\theta_0 \sim \frac{1}{\sqrt{d}}\mathsf{N}(0, I)$. Clearly, as ε grows there is a wider range of Pareto-optimal estimators and the two measures of risks would deviate further from each other. When ε becomes smaller, the two measures of standard and adversarial risks get closer to each other and so the Pareto-optimal curve becomes shorter.

3.2. Binary classification

We next consider the problem of binary classification under a Gaussian mixture data model. Under this model, each data point belongs to one of two two classes $\{\pm 1\}$ with corresponding probabilities π_+, and $\pi_- = 1 - \pi_+$. The feature vectors in each class are generated independently according to an isometric Gaussian distribution with mean $\{\pm\mu\}$ depending on the class. In other words, given label $y_i \in \{\pm 1\}$, the feature vector $x_i \in \mathbb{R}^d$ is drawn from $\mathsf{N}(y_i\mu, \Sigma)$.

We focus on class of linear classifiers $\{x^\mathsf{T}\theta : \theta \in \mathbb{R}^d\}$. Given a model θ the predicted labels are simply given as $\mathrm{sign}(x^\mathsf{T}\theta)$. We consider 0-1 loss $\ell(\theta; z) = \mathbb{I}(\hat{y} \neq y) = \mathbb{I}(yx^\mathsf{T}\theta \leq 0)$. We also consider Wasserstein adversarial training with distance

$$d(z, \tilde{z}) = \|x - \tilde{x}\|_{\ell_r} + \infty \cdot \mathbb{I}\{y \neq \tilde{y}\}\,, \tag{22}$$

for $z = (x, y)$ and $\tilde{z} = (\tilde{x}, \tilde{y})$. Our next results is on characterizing the standard risk and the Wasserstein adversarial risk for this model.

Proposition 3.5. *Consider binary classification with Gaussian mixture data model and 0-1 loss. Let* $a_\theta := \frac{\mu^\mathsf{T}\theta}{\|\Sigma^{1/2}\theta\|_{\ell_2}}$. *Then, for a linear classifier* $x \mapsto \mathrm{sgn}(x^\mathsf{T}\theta)$, *the standard risk is given by*

$$\mathsf{SR}(\theta) = \Phi(-a_\theta)\,,$$

where $\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{t^2}{2}}\mathrm{d}t$ *denotes the c.d.f of a standard Gaussian distribution. In addition, define the function* $F(\theta, \gamma) : \mathbb{R}^{d+1} \mapsto \mathbb{R}_{\geq 0}$ *as follows:*

$$\begin{aligned} F(\theta,\gamma) =& \frac{\gamma}{b_\theta}\varepsilon^2 + \Phi\Big(\sqrt{\frac{2}{\gamma}} - a_\theta\Big) \\ &+ \frac{\gamma}{2}\Big\{\Big(a_\theta + \sqrt{\frac{2}{\gamma}}\Big)\varphi\Big(a_\theta - \sqrt{\frac{2}{\gamma}}\Big) - a_\theta\varphi(a_\theta) \\ &+ (a_\theta^2 + 1)\Big[\Phi\Big(a_\theta - \sqrt{\frac{2}{\gamma}}\Big) - \Phi(a_\theta)\Big]\Big\}\,, \end{aligned} \tag{23}$$

with $b_\theta := \frac{\|\Sigma^{1/2}\theta\|_{\ell_2}^2}{\|\theta\|_{\ell_q}^2}$, ℓ_q *denoting the dual norm of* ℓ_r *(i.e.,* $\frac{1}{r} + \frac{1}{q} = 1$*) and* $\varphi(t) := \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$ *standing for the p.d.f*

of a standard Gaussian distribution. Then, the Wasserstein adversarial risk with $p = 2$ and metric $d(\cdot, \cdot)$ given by (22) is characterized as

$$\mathsf{AR}(\theta) = \inf_{\gamma \geq 0} F(\theta, \gamma)\,. \tag{24}$$

Note that as an implication of Proposition 3.5, the standard risk $\mathsf{SR}(\theta)$ and the adversarial risk $\mathsf{AR}(\theta)$ depend on the estimator θ only through the components a_θ and b_θ.

We next characterize the Pareto optimal front for the region $\{(\mathsf{SR}(\theta), \mathsf{AR}(\theta)) : \theta \in \mathbb{R}^d\}$. Since the 0-1 loss $\mathbb{I}(yx^\mathsf{T}\theta \leq 0)$ is convex in θ, both the standard risk and the adversarial risks are convex functions of θ (by a similar argument given prior to Theorem 3.3.)

Theorem 3.6. *Assume the setting of Proposition 3.5 and consider the following minimization problem*

$$(\theta_*^\lambda, \gamma_*^\lambda) := \arg\min_{\gamma \geq 0, \theta} \lambda\Phi(-a_\theta) + F(\theta, \gamma)\,. \tag{25}$$

The Pareto optimal curve is given by

$$\{(\Phi(-a_{\theta_*^\lambda}), F(\theta_*^\lambda, \gamma_*^\lambda)) : \lambda \geq 0\}\,.$$

Theorem 3.6 follows from the fact that the Pareto front of a convex set is characterized by intersection points of the set with the supporting hyperplanes.

Remark 3.7. *For $r = q = 2$ and $\Sigma = I$, we have $b_\theta = 1$. In this case the objective of (25) is decreasing in a_θ and since $|a_\theta| \leq \|\mu\|_{\ell_2}$, it is minimized at $a_\theta = \|\mu\|_{\ell_2}$. In addition, $\mathsf{SR}(\theta)$ is decreasing in a_θ and is minimized at the same value of $a_\theta = \|\mu\|_{\ell_2}$. Therefore, by introducing $c := \|\mu\|_{\ell_2}$, the Pareto-optimal curve shrinks to a single point given by*

$$\mathsf{SR} = \Phi(-c)\,, \tag{26}$$

$$\mathsf{AR} = \inf_{\gamma \geq 0} \Big[\gamma\varepsilon^2 + \Phi\Big(\sqrt{\frac{2}{\gamma}} - c\Big) + \frac{\gamma}{2}\Big\{\Big(c + \sqrt{\frac{2}{\gamma}}\Big)\varphi\Big(c - \sqrt{\frac{2}{\gamma}}\Big) - c\varphi(c) + (c^2+1)\Big[\Phi\Big(c - \sqrt{\frac{2}{\gamma}}\Big) - \Phi(c)\Big]\Big\}\Big].$$

In other words, the tradeoff between standard and adversarial risks, achieved by linear classifiers, vanishes in this case and the estimators in direction of the class average μ are optimal with respect to both standard risk and the Wasserstein adversarial risks.

We refer to Section A.5 for the proof of Remark 3.7.

Figure 2 showcases the effect of different factors in a binary classification setting on the Pareto-optimal tradeoff between standard and adversarial risks. Here the features x are drawn from $\mathsf{N}(y\mu, \Sigma)$, with $\Sigma_{ij} = \rho^{|i-j|}$. The class average μ has i.i.d entries from $\mathsf{N}(0, 1/d)$ with $d = 10$. In Figure 2a, we investigate the role of the norm r used in the Wasserstein adversary model, cf. Equation (22). As discussed in Remark 3.7, when $r = 2$, the tradeoff between standard and adversarial risks vanishes and the estimators in direction of the class average μ are optimal with respect to both standard risk and the Wasserstein adversarial risks.

Figure 2b illustrates the effect of dependency among features ρ on optimal tradeoff between standard and adversarial risks. In this setting $r = \infty$ and $\varepsilon = 0.3$. From the result of Theorem 3.6, we see that these risks very much depend on the interaction between the class average μ and the features covariance Σ and so the curves are shifted in highly nontrivial way depends on the value of ρ when we fix μ.

The role of adversary's budget ε is depicted in figure 2c in which $r = \infty$, $\rho = 0$. Similar to the linear regression setting, when ε is small the two measures of risk are close to each other and we have a small range of Pareto-optimal models. As ε grows, the standard risk and the adversarial risks differ significantly and we get a wide range of Pareto-optimal models.

3.3. Learning nonlinear functions

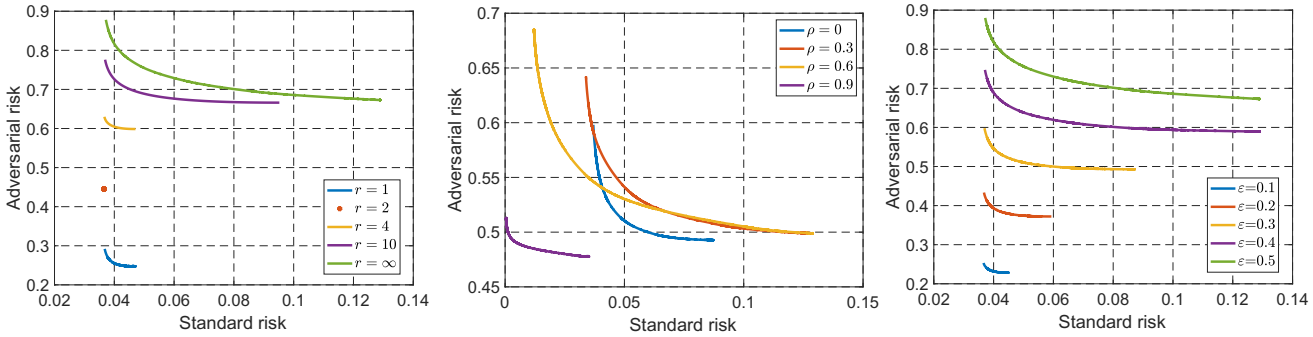
We next investigate the tradeoff between standard and adversarial risk for the problem of learning an unknown function over the d-dimensional sphere $\mathbb{S}^{d-1}$. More precisely, we consider the following data generative model:

$$y = f_d(x) + w\,, \tag{27}$$

with $x \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, the d-dimensional sphere of radius $\sqrt{d}$, and $w \sim \mathsf{N}(0, \sigma^2)$ independent of x. We consider fitting a random features model to data generated according to (27). The class of random features model is given by

$$\mathcal{F}_{\mathrm{RF}}(\theta, U) = \Big\{ f(x, \theta, U) := \sum_{i=1}^N \theta_i \sigma(u_i^\mathsf{T} x)\,, \quad \text{with } \theta_i \in \mathbb{R},\ i = 1, \dots, N \Big\}, \tag{28}$$

where $U \in \mathbb{R}^{N \times d}$ is a matrix whose i-th row is the vector u_i, uniformly drawn from $\mathbb{S}^{d-1}(1)$, independently from data. The random features model can be equivalently represented by two-layer neural network with the first-layer weights U chosen randomly and $\theta = (\theta_i)_{1 \leq i \leq N}$ corresponding to the second-layer weights. The random features model was introduced by (Rahimi & Recht, 2007) for scaling kernel methods to large datasets. There is indeed a substantial literature drawing connections between random features models, kernel methods and fully trained neural networks (Daniely



(a) Pareto optimal curve for several ℓ_r norms on feature space with $d = 10$, $\varepsilon = 0.5$ and $\rho = 0$.

(b) Pareto optimal curve for several feature dependency values (ρ) with $d = 10$, $\varepsilon = 0.3$, and $r = \infty$.

(c) Pareto optimal curve for several adversary's budget ε with $d = 10$, $r = \infty$, and $\rho = 0$.

Figure 2. The effect of defined ℓ_r norm on feature space, dependency across features (ρ) and adversary's power ε on Pareto optimal tradeoff between adversarial and standard risks in binary classification under Gaussian mixture model.

et al., 2016; Daniely, 2017; Jacot et al., 2018; Li & Liang, 2018). In (Mei & Montanari, 2019), the generalization error (standard risk) of random features model was precisely characterized for the problem of learning a function $f_d(\cdot)$ over $\mathbb{S}^{d-1}(\sqrt{d})$ in the regime where the network width N, sample size n and feature dimension d grow in proportion. The nonlinear model considered in (Mei & Montanari, 2019) is of the form

$$f_d(x) = \beta_{d,0} + x^{\mathsf{T}}\beta_{d,1} + f_d^{\mathrm{NL}}(x)\,, \tag{29}$$

with the nonlinear component $f_d^{\mathrm{NL}}(x)$ is a centered isotropic Gaussian process indexed by x. We follow the same model and consider the following random quadratic function

$$f_d(x) = \beta_{d,0} + x^{\mathsf{T}}\beta_{d,1} + \frac{F_*}{d}[x^{\mathsf{T}}Gx - \mathrm{tr}(G)]\,, \tag{30}$$

for some fixed $F_* \in \mathbb{R}$ and $G \in \mathbb{R}^{d\times d}$ a random matrix with i.i.d entries from $\mathsf{N}(0,1)$.

Our goal is to study the Pareto-optimal tradeoff between standard and adversarial risks for this learning setting, achieved by the class of random features model (28). The standard risk in this setting is given by

$$\begin{aligned}\mathsf{SR}(\theta) &= \mathbb{E}_{x,y}\left[(y - \theta^{\mathsf{T}}\sigma(Ux))^2\right] \\ &= \mathbb{E}_x\left[(f_d(x) - \theta^{\mathsf{T}}\sigma(Ux))^2\right] + \sigma^2\,.\end{aligned} \tag{31}$$

For the Wasserstein adversarial risk we use the following corollary which is obtained by specializing Proposition 2.3 to random features model.

Corollary 3.8. *Consider the class of feature model given by (28). In this case, the 2-Wasserstein adversarial risk with distance $d(\cdot,\cdot)$ (12) admits the following first-order approximation:*

$$\begin{aligned}\mathsf{AR}(\theta) &= \mathsf{SR}(\theta) \\ &\quad + 2\varepsilon\,\mathbb{E}_x\Big[\left[(f_d(x) - \theta^{\mathsf{T}}\sigma(Ux))^2 + \sigma^2\right] \\ &\qquad \times \left\|U^{\mathsf{T}}\mathrm{diag}(\sigma'(Ux))\theta\right\|_{\ell_2}^2\Big]^{1/2} + O(\varepsilon^2)\,,\end{aligned} \tag{32}$$

with $\sigma'(\cdot)$ denoting the derivative of the activation $\sigma(\cdot)$ and $\mathsf{SR}(\theta)$ given by (31).

The proof of Corollary 3.8 is given in Appendix A.6. The standard risk is quadratic and hence convex in θ. The adversarial risk is also convex in θ (it follows from the fact that pointwise maximization preserves convexity.) Therefore, for small values of ε (weak adversary) the first order approximation of $\mathsf{AR}(\theta)$ is also convex in θ. As such, (almost) all Pareto optimal points are given by minimizing a weighted combination of the two risk measures as the weight λ varies in $[0,\infty)$. Namely,

$$\theta_\lambda := \arg\min_\theta\ \lambda\mathsf{SR}(\theta) + \mathsf{AR}(\theta)\,, \tag{33}$$

with $\mathsf{SR}(\theta)$ given by (31), and $\mathsf{AR}(\theta)$ given by (32).

We use the above characterization to derive the Pareto-optimal tradeoff curves between standard and adversarial risks for learning function $f_d(x)$, given by (30) with $F_* = 1$, $\beta_{d,0} = 0$, and $\beta_{d,1} \in \mathbb{R}^d$ with i.i.d entries $\sim \mathsf{N}(0, 1/d)$. The data are generated according to (27) with $\sigma = 2$, $d = 10$ and $N \in \{250, 500, 750, 1000\}$. To compute θ_λ we use empirical loss with $n = 500K$ samples of $x \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. For each value of λ and N we generate 15 realization of weights U and compute θ_λ for each realizations using gradient descent on the loss function (33). The Pareto optimal points $\{\mathsf{SR}(\theta_\lambda), \mathsf{AR}(\theta_\lambda) : \lambda \geq 0\}$ are plotted in Figure 3

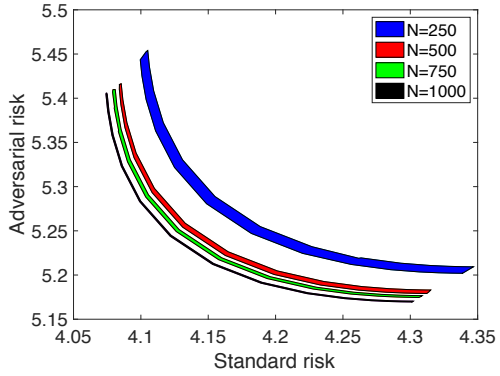


Figure 3. Pareto-optimal tradeoff curves for learning random quadratic functions using random features model. Data is generated according to (27) with $\sigma = 2$ and $f_d(x)$ given by (30). Here, $d = 10$ and N is the number of random features (width of the neural network).

for $\varepsilon = 0.2$. As we see for each value of N the tradeoff curves concentrate as N grows implying that the estimator θ_λ becomes independent of the specific realization of weights U. Also we observe that the tradeoff between standard and adversarial risks exist even for large values of N. Interestingly, as the network width N grows both the standard risk and adversarial risk decrease but the tradeoff between them clearly remains (the length of Pareto front does not shrink).

4. Conclusion

Linear regression and binary classification are two simple, yet foundational settings in machine learning and still the full effect of adversarial training is not known for these settings. In this work we focus on distribution perturbing adversary and provide a framework on how to think about the tradeoff between the Standard risk (SR) and the Adversarial risk (AR), its existence and its quantitative behavior with respect to data distribution and the hypotheses class. Note that these are non-trivial questions and previously there has been specific "examples" to hint such tradeoff. A tantalizing question is whether one can remove this tradeoff (or improve SR and AR simultaneously) by considering a more complex class of hypotheses. Our discussion in Section 3.3 is a first attempt to answer this question for random features model.

Acknowledgements

A. Javanmard is partially supported by an Adobe Data Science Faculty Research Award, Sloan Research Fellowship, and the NSF CAREER Award DMS-18444. The authors would like to thank the anonymous reviewers for their thoughtful comments that helped us to improve our work.

References

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420, International Conference on Machine Learning*, 2018. 1

Bartl, D., Drapeau, S., Obloj, J., and Wiesel, J. Robust uncertainty sensitivity analysis. *arXiv preprint arXiv:2006.12022*, 2020. 4

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Srndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013. 1

Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 3–14, 2017. 1

Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., Wagner, D., and Zhou, W. Hidden voice commands. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pp. 513–530, 2016. 1

Chen, L., Ye, Y., and Bourlai, T. Adversarial machine learning in malware detection: Arms race between evasion attack and defense. In *2017 European Intelligence and Security Informatics Conference (EISIC)*, pp. 99–106. IEEE, 2017. 1

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. *arXiv preprint arXiv:1704.08847, International Conference on Machine Learning*, 2017. 1

Daniely, A. Sgd learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, pp. 2422–2430, 2017. 8

Daniely, A., Frostig, R., and Singer, Y. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances In Neural Information Processing Systems*, pp. 2253–2261, 2016. 7

Dobriban, E., Hassani, H., Hong, D., and Robey, A. Provable tradeoffs in adversarially robust classification. *arXiv preprint arXiv:2006.05161*, 2020. 2

Dong, Y., Deng, Z., Pang, T., Su, H., and Zhu, J. Adversarial distributional training for robust deep learning. *arXiv preprint arXiv:2002.05999*, 2020. 3

Gao, R. and Kleywegt, A. J. Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016. 4

Gao, R., Chen, X., and Kleywegt, A. J. Wasserstein distributionally robust optimization and variation regularization. *arXiv preprint arXiv:1712.06050v3*, 2020. 4

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015a. 1

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572, International Conference on Learning Representations*, 2015b. 1

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018. 8

Javanmard, A. and Soltanolkotabi, M. Precise statistical analysis of classification accuracies for adversarial training. *arXiv preprint arXiv:2010.11213*, 2020. 2

Javanmard, A., Soltanolkotabi, M., and Hassani, H. Precise tradeoffs in adversarial training for linear regression. volume 125 of *Proceedings of Machine Learning Research, Conference of Learning Theory (COLT)*, pp. 2034–2078. PMLR, 09–12 Jul 2020. 2

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 1

Li, Y. and Liang, Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pp. 8157–8166, 2018. 8

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018a. 1

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083, International Conference on Learning Representations*, 2018b. 2

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019. 8

Papernot, N., McDaniel, P., Swami, A., and Harang, R. Crafting adversarial input sequences for recurrent neural networks. In *MILCOM 2016-2016 IEEE Military Communications Conference*, pp. 49–54. IEEE, 2016a. 1

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. IEEE, 2016b. 1

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017. 1

Pydi, M. S. and Jog, V. Adversarial risk via optimal transport and optimal couplings. In *International Conference on Machine Learning*, pp. 7814–7823. PMLR, 2020. 3, 4

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. 1

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J. C., and Liang, P. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019. 2

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20:1177–1184, 2007. 7

Staib, M. and Jegelka, S. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, 2017. 3, 4

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *International Conference on Learning Representations (ICLR)*, 2014. 1

Taheri, H., Pedarsani, R., and Thrampoulidis, C. Asymptotic behavior of adversarial training in binary classification. *arXiv preprint arXiv:2010.13275*, 2020. 2

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018. 2

Vaidya, T., Zhang, Y., Sherr, M., and Shields, C. Cocaine noodles: exploiting the gap between human and machine speech recognition. In *9th {USENIX} Workshop on Offensive Technologies ({WOOT} 15)*, 2015. 1

Wong, E. and Kolter, J. Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 5283–5292, 2018. 1

Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., and Xu, W. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 103–117, 2017. 1