Asymptotic Behavior of Adversarial Training in Binary Linear Classification

Hossein Taheri
Electrical and Computer Engineering
University of California, Santa Barbara
Santa Barbara, USA
hossein@ucsb.edu

Ramtin Pedarsani
Electrical and Computer Engineering
University of California, Santa Barbara
Santa Barbara, USA
ramtin@ucsb.edu

Christos Thrampoulidis
Electrical and Computer Engineering
University of British Columbia
Vancouver, Canada
cthrampo@ece.ubc.ca

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834717

***Abstract*—Adversarial training using empirical risk minimization is the state-of-the-art method for defense against adversarial attacks, that is against small additive adversarial perturbations applied to test data leading to misclassification. Despite being successful in practice, understanding generalization properties of adversarial training in classification remains widely open. In this paper, we take the first step in this direction by precisely characterizing the robustness of adversarial training in binary linear classification. Specifically, we consider the high-dimensional regime where the model dimension grows with the size of the training set at a constant ratio. Our results provide exact asymptotics for both standard and adversarial test errors under ℓ_∞-norm bounded perturbations in a generative Gaussian-mixture model. We use our sharp error formulae to explain how the adversarial and standard errors depend upon the over-parameterization ratio, the data model, and the attack budget. Finally, by comparing with the robust Bayes estimator, our sharp asymptotics allow us to study fundamental limits of adversarial training.**

I. Introduction

Several machine learning models ranging from simple linear classifiers to complex deep neural networks have been shown to be prone to adversarial attacks, i.e., small additive perturbations to the data that cause the model to predict a wrong label [SZS+13], [MDFF16]. The requirement for robustness against adversaries is crucial for the safety of systems that rely on decisions made by these algorithms (e.g., in self-driving cars). With this motivation, over the past few years, there have been remarkable efforts by the research community to construct defense mechanisms, e.g., see [SN20], [CAD+18] for a survey. Among many proposals in the already rich literature, perhaps the most popular approach is that of adversarial training [GSS14]. Among many favorable properties, adversarial training is flexible and easy-to-adjust to different types of data perturbations and has also been shown to achieve state-of-the-art performance in several tasks [MMS+17]. However, despite major recent progress in the study and implementation of adversarial training, its efficacy has been mainly shown empirically without providing much theoretical understanding. Indeed, many questions regarding its theoretical properties remain open even for simple models. For instance, how does the adversarial/standard error depend on the adversary's budget during training time and test time? How do they depend on the over-parameterization ratio? What is the role of the chosen loss function?

In this paper, we consider the adversarial training problem for ℓ_∞-norm bounded perturbations in classification tasks, which solves the following robust empirical risk minimization (ERM) problem:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^l} \sum_{i=1}^{m} \max_{\|\boldsymbol{\delta}_i\|_\infty \le \varepsilon_{\rm tr}} \widetilde{\mathcal{L}}\left(y_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i+\boldsymbol{\delta}_i)\right) + r\|\boldsymbol{\theta}\|_2^2. \tag{1}$$

Here, $\{(\boldsymbol{x}_i, y_i)\}_{i\in[m]} \in \mathbb{R}^n \times \{\pm 1\}$ is the training set, $\boldsymbol{\delta}_i \in \mathbb{R}^n$ are the perturbations with l the dimension of the feature space, $f_{\boldsymbol{\theta}} : \mathbb{R}^n \to \mathbb{R}$ is a model parameterized by a vector $\boldsymbol{\theta} \in \mathbb{R}^l$, $\varepsilon_{\rm tr}$ is a user-specified tunable parameter that can be interpreted as the adversary's budget during training, and r is the ridge-regularization parameter. Once the robust classifier $\widehat{\boldsymbol{\theta}}$ is obtained by (1), the *adversarial error / robust classification error* is given by $\mathbb{E}_{\boldsymbol{x},y}[\max_{\|\boldsymbol{\delta}\|_\infty\le\varepsilon_{\rm ts}} \mathbf{1}_{\{y\neq \mathtt{sign}(f_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x}+\boldsymbol{\delta}))\}}]$, where $\mathbf{1}_{\{\cdot\}}$ is the $0/1$-indicator function, $(\boldsymbol{x}, y) \in \mathbb{R}^n \times \{\pm 1\}$ is a test sample drawn from the same distribution as that of the training dataset, $\varepsilon_{\rm ts}$ is the budget of the adversary, and $f_{\widehat{\boldsymbol{\theta}}}$ uses the trained parameters $\widehat{\boldsymbol{\theta}}$ and the fresh sample $\boldsymbol{x}$ to output a label guess. The standard classification error is given by the same formula by simply setting $\varepsilon_{\rm ts} = 0$.

The goal of this paper is to precisely analyze the performance of adversarial training in (1) for binary classification with linear models i.e., $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \boldsymbol{x}\rangle$. In our proof we use the Convex-Gaussian-Min-max-Theorem (CGMT) [TOH15], [Sto09], [Sto13] and in particular its applications to the convex ERM that enables its precise analysis, e.g., [TAH18], [MRSY19], [SAH19], [TPT20b], [TPT21]. However, compared to previous works, we develop a new analysis for robust optimization.

Our main contributions are summarized as follows:

- We precisely analyze, for the first time, the performance of adversarial training with ℓ_∞ attacks in binary classification for

The authors acknowledge support by NSF grants 1909320, 2003035, 193464, 2009030 and a GR8 award from KAUST.

the Gaussian Mixtures Model. See Section III.

• Numerical illustrations in Section III-A show tight agreements between our theoretical and empirical results and also allow us to draw intriguing conclusions regarding the behavior of adversarial and standard errors as functions of key problem parameters such as the sampling ratio $\delta := m/n$, the budget of the adversary $\varepsilon_{\rm ts}$, and the robust-optimization hyper-parameter $\varepsilon_{\rm tr}$ in our studied settings.

A. Novelty and Prior Works

Relevant to the flavour of our results, the recent work [JSH20] studies precise tradeoffs and performance analysis in adversarial training with linear regression with ℓ_2 perturbations and isotropic Gaussian data. Compared to [JSH20], our results hold for binary models. Moreover, we consider regularized ERM allowing us to study the behavior of adversarial training in the over-parameterized regime in the limit of $\lambda \to 0$. Similar results on the behavior of adversarial training in classification are only derived in a *concurrent* work by [JS20]. On the one hand, compared to [JS20] our analysis applies to the *regularized* ERM. Additionally, we examine how our formulae on adversarial training compare with those of the Bayes robust estimator. On the other hand, [JS20] extend their analysis to robust support vector machines (SVM). Note however that we can retrieve the same results regarding the performance of adversarially-robust SVM by evaluating our formulae on regularized ERM with logistic loss and vanishing regularization parameter.

To see, at a high-level, why adversarial training differs from standard ERM or standard SVM analysis note the following complications in the analysis. First, because adversarial training is formulated as a min-max optimization, it is not at all apparent that the machinery of Gaussian comparison theorems applies. Second, the performance metric here is robust error (rather than standard error), and we show that this changes the statistics that needs to be tracked by the CGMT analysis. Third, the primary optimization to which we eventually apply the CGMT involves an "effective" ℓ_1-regularizer which unlike previous works appears inside the argument of the loss function, requiring new techniques to scalarize the auxiliary optimization.

The Adversarial Bayes risk for Gaussian-mixtures has been recently characterized in [BCM19]. Here, we combine their results with our precise asymptotics on the practically relevant adversarial training method, allowing us to investigate fundamental limits of adversarial training. The references [CRWP19], [AZL20] discuss optimization landscape of adversarial training, however these works do not address generalization properties of adversarial training, as done in this paper. Another related line of work studies trade-offs between the standard and adversarial errors e.g., see [TSE+18], [RXY+19], [ZYJ+19], [DHHR20], but for simpler algorithms and data models, rather than adversarial training, which we focus on here. The benefits of unlabeled data in robustness have been investigated in several works, e.g. [RXY+20], [CRS+19]. An exciting direction opening up with our analysis is investigating adversarial training performance for random features and neural tangent models. To date, precise asymptotics for such models have been obtained only very recently and for the simpler problem of standard ERM [MM19], [GMKZ20], [DL20], [DL21], [GMMM19].

II. Problem Formulation

In this section, we describe the data model, the specific form of (1), and the asymptotic regime for which our results hold. After this section, it is understood that all our results hold in the setting described here without any further explicit reference.

A. Data Model

We study Gaussian Mixture model (GMM) where the conditional distribution of the feature vectors is a Gaussian with mean $\pm\boldsymbol{\theta}_n^\star$ (depending on the label $y_i \in \{\pm 1\}$). The subscript n emphasizes the dependence on dimension. Formally, the GMM assumes

$$\mathbb{P}(y_i = 1) = \pi \in [0,1], \quad \boldsymbol{x}_i | y_i \sim \mathcal{N}(y_i \boldsymbol{\theta}_n^\star, \mathbb{I}_n). \tag{2}$$

We assume that each entry of the true vector $\boldsymbol{\theta}^\star$ is sampled iid from a fixed distribution $\mathcal{D}$, i.e., $\boldsymbol{\theta}_i^\star \overset{\text{iid}}{\sim} \mathcal{D}$. Moreover, without loss of generality we assume that $\boldsymbol{\theta}^\star$ is normalized such that $\|\boldsymbol{\theta}^\star\|_2 = 1$.

B. Asymptotic Regime

We consider the high-dimensional asymptotic regime in which the size m of the training set and the dimension n of the feature space grow large at a proportional rate. Formally, $m, n \to \infty$ at a fixed ratio $\delta = m/n$.

C. Robust Learning

Let $\widehat{\boldsymbol{\theta}}_n$ be a linear classifier trained on data generated according to the data model (2). As is typical, given $\widehat{\boldsymbol{\theta}}_n$, a decision is made about the label of $\boldsymbol{x}$ based on $\texttt{sign}(\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_n \rangle)$. Thus, letting y be the label of a fresh sample $\boldsymbol{x}$, the *standard error* is given by

$$\mathcal{E}(\widehat{\boldsymbol{\theta}}_n) \triangleq \mathbb{E}_{\boldsymbol{x},y}\left[\mathbf{1}_{\left\{y \neq \texttt{sign}\left(\langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}}_n \rangle\right)\right\}}\right]. \tag{3}$$

Here, the expectation is over a fresh pair $(\boldsymbol{x}, y)$ also generated according to the GMM model. Next, we define the adversarial error with respect to a worst-case ℓ_∞-norm bounded additive perturbation. Let $\varepsilon_{\rm ts} \geq 0$ be the budget of the adversary. Then, the *adversarial error* is defined as follows:

$$\mathcal{E}_{\varepsilon_{\rm ts}}(\widehat{\boldsymbol{\theta}}_n) \triangleq \mathbb{E}_{\boldsymbol{x},y}\left[\max_{\|\boldsymbol{\delta}\|_\infty \leq \varepsilon_{\rm ts}} \mathbf{1}_{\left\{y \neq \texttt{sign}\left(\langle \boldsymbol{x}+\boldsymbol{\delta}, \widehat{\boldsymbol{\theta}}_n \rangle\right)\right\}}\right]. \tag{4}$$

Adversarial training leads to a classifier $\widehat{\boldsymbol{\theta}}_n$ that solves the robust optimization problem (1) with $\widetilde{\mathcal{L}}(y, f_{\boldsymbol{\theta}}(\boldsymbol{x}+\boldsymbol{\delta}))$ replaced by $\mathcal{L}(y\langle \boldsymbol{\theta}, \boldsymbol{x}+\boldsymbol{\delta}\rangle)$. The loss function $\mathcal{L} : \mathbb{R} \to [0,\infty)$ is chosen as a convex approximation to the 0/1 loss. Specifically, throughout the paper, we assume that $\mathcal{L}$ is convex and

decreasing. This includes popular choices such as the logistic, hinge and exponential losses.

III. Main Results

In this section, we focus on the case of bounded ℓ_∞-perturbations. Specifically, let $\widehat{\boldsymbol{\theta}}_n$ be a solution to the following robust minimization:

$$\min_{\boldsymbol{\theta}_n} \sum_{i=1}^{m} \max_{\|\boldsymbol{\delta}_i\|_\infty \le \frac{\varepsilon_{\rm tr}}{\sqrt{n}}} \mathcal{L}\left(y_i \left\langle \boldsymbol{x}_i + \boldsymbol{\delta}_i, \boldsymbol{\theta}_n \right\rangle\right) + r\|\boldsymbol{\theta}_n\|_2^2. \quad (5)$$

In our asymptotic setting, $\varepsilon_{\rm tr}$ is of constant order and the factor $1/\sqrt{n}$ in front of it is the proper normalization needed to ensure that the perturbations norm $\|\boldsymbol{\delta}_i\|_2$, in comparable to the norm of the true vector $\|\boldsymbol{\theta}_n^\star\|_2$, i.e., both are constant in the high-dimensional limit $n \to \infty$. We explain this normalization further in Section III-B.

Before presenting our main result, we need to introduce some necessary definitions. We write

$$\mathcal{M}_f(x;\kappa) \triangleq \min_v \frac{1}{2\kappa}(x-v)^2 + f(v), \quad (6)$$

for the Moreau envelope of a function $f:\mathbb{R}\to\mathbb{R}$ at $x\in\mathbb{R}$ with parameter $\kappa>0$ [RW09]. We also define the following min-max optimization over eight scalar variables. Denote $\bar{\mathbf{v}} \triangleq (\alpha,\tau_1,w,\mu,\tau_2,\beta,\gamma,\eta)$ and define $f:\mathbb{R}^8\to\mathbb{R}$ as follows:

$$f(\bar{\mathbf{v}}) \triangleq -\gamma w - \frac{\mu^2\tau_2}{2\alpha} - \frac{\alpha\beta^2}{2\delta\tau_2} - \frac{\alpha\tau_2}{2} + \frac{\beta\tau_1}{2} + \eta\mu - \frac{\eta^2\alpha}{2\tau_2},$$

We introduce the following min-max objective based on the eight scalars,

$$\min_{\substack{\alpha,\tau_1,w\in\mathbb{R}_+,\\ \mu\in\mathbb{R}}} \max_{\substack{\tau_2,\beta,\gamma\in\mathbb{R}_+,\\ \eta\in\mathbb{R}}} \mathbb{E}\left[\mathcal{M}_{\mathcal{L}}\left(\sqrt{\mu^2+\alpha^2}\,G + \mu - w; \tau_1/\beta\right)\right]$$
$$+\ \gamma\varepsilon_{\rm tr}\,\mathbb{E}\left[\mathcal{M}_{\ell_1}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta}}H + \frac{\alpha\eta}{\tau_2}Z; \frac{\alpha\gamma\varepsilon_{\rm tr}}{\tau_2}\right)\right] + f(\bar{\mathbf{v}}), \quad (7)$$

where $G, H \overset{\rm iid}{\sim} \mathcal{N}(0,1)$ and $Z\sim\mathcal{D}$. Notice that the objective function of (7) depends explicitly on the sampling ratio δ and on the training parameter $\varepsilon_{\rm tr}$. Moreover, it depends implicitly on $\boldsymbol{\theta}_n^\star$ via $\mathcal{D}$, and on the specific loss $\mathcal{L}$ via its Moreau envelope. The nature of allowed perturbations (the ℓ_∞-type) is reflected in (7), via the Moreau-envelope of the dual-norm (the ℓ_1 norm).

We are now ready to state our main result in Theorem 1, which establishes a relation between the solutions of (7) and the adversarial risk of the robust classifier $\widehat{\boldsymbol{\theta}}_n$. The proof is deferred to the long version of the paper [TPT20a].

Theorem 1. *Assume that the training dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, is generated according to the data model* (2)*. Consider the robust classifiers $\{\widehat{\boldsymbol{\theta}}_n\}$, obtained by adversarial training in* (5)*. Then, the high-dimensional limit for the adversarial error satisfies,*

$$\mathcal{E}_{\frac{\varepsilon_{\rm ts}}{\sqrt{n}}}(\widehat{\boldsymbol{\theta}}_n) \xrightarrow{P} Q\Big(\frac{\mu^\star - w^\star\,\varepsilon_{\rm ts}/\varepsilon_{\rm tr}}{\sqrt{\mu^{\star 2}+\alpha^{\star 2}}}\Big). \quad (8)$$

where $Q(\cdot)$ denotes the Gaussian Q-function and $(\alpha^\star,\mu^\star,w^\star)$ is the unique solution to the scalar minimax problem (7)*.*

The asymptotics for adversarial error in Theorem 1 are precise in the sense that they hold with probability 1, as $m,n\to\infty$. In the following section, we demonstrate the precise theoretical values and the corresponding numerical values.

A. Numerical Illustrations

In this section, we illustrate the theoretical predictions for various values of the different problem parameters, including $\delta = m/n$ and the attack budgets $\varepsilon_{\rm tr}$ and $\varepsilon_{\rm ts}$. For numerical results here, we focus on the hinge-loss i.e., $\mathcal{L}(t) = \max(1-t,0)$ and on the GMM with isotropic features. We further assume that $\mathcal{D}$ is standard normal and fix regularization parameter $r = 10^{-4}$. To solve (7), we derive the solution of the corresponding saddle-point equations by iterating over the equations and finding the fixed-point solution after 100 iterations. For the numerical results, we set $n = 200$ and solve the ERM problem (5) by gradient descent. The resulting estimator is used to compute the adversarial test error by evaluating (3) on a test set of 3×10^3 samples. We then average the results over 20 independent experiments. The results for both numerical and theoretical values are depicted in Figures 1-2. Next, we discuss some of the insights obtained from these figures.

a) Impact of $\boldsymbol{\delta}$ on standard/adversarial errors.: Figure 1 depicts the adversarial and standard errors as a function of $\delta = m/n$. The dashed lines show the *Bayes Adversarial Error*, i.e., the smallest adversarial error obtained by any classifier [BCM19], [DWR20], [DHHR20]. Note that both errors decrease as the sampling ratio δ grows, with the adversarial error approaching the Bayes adversarial error of the corresponding value of $\varepsilon_{\rm ts}$. More generally, in light of comparison between the error formulae of Theorem 1 and the Bayes adversarial error, Figure 1 provides a means to quantify the sub-optimality gap of adversarial training for all values of the oversampling ratio $\delta>0$ and for different values of the adversary's budget. A related study was performed in [SST+18], but therein the authors derive error bounds for a simple averaging estimator. Instead, our analysis is precise and holds for the broader case of convex decreasing losses. Next, we comment on the shape of the error curves as a function of the sampling ratio. Note that a second sharp decrease in standard and adversarial errors appears right after an separability threshold $\delta^*_{\frac{\varepsilon_{\rm tr}}{\sqrt{n}}}$, which we define as the maximum value of δ for which the data-points are $(\ell_\infty, \frac{\varepsilon_{\rm tr}}{\sqrt{n}})$-separable (for definition, see the discussion on Robust Separability in Section IV). This constantly decreasing behavior of the error is in contrast to the corresponding behavior in linear regression with ℓ_2 perturbations and ℓ_2 loss analyzed in [JSH20], where error based on δ starts rising after the first decrease and then again decreases as δ grows. This double-descent behavior can be considered as extension of the more familiar double-descent behavior in standard ERM (first observed in numerous high-dimensional machine learning models [BHMM18], [BHX19],

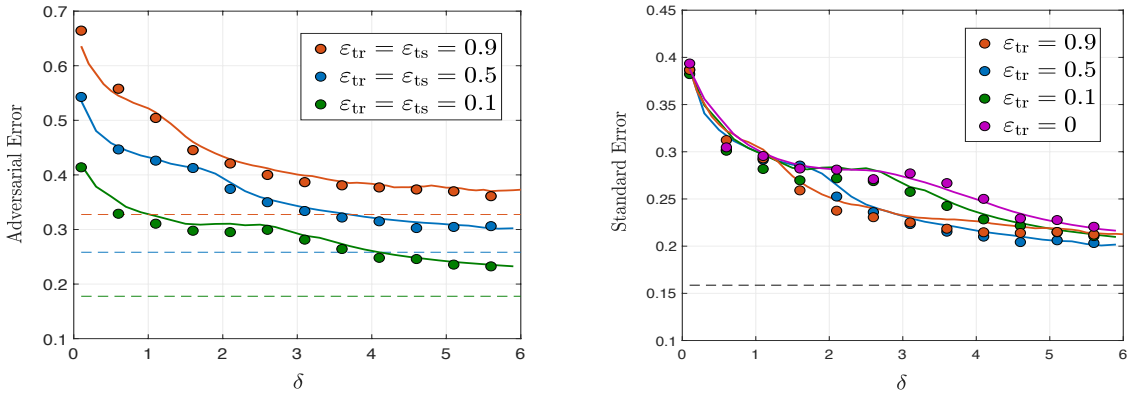

Fig. 1: Adversarial/Standard test error based on $\delta := m/n$. Solid lines correspond to theoretical predictions while markers denote the empirical results derived by solving ERM using gradient descent ($r = 10^{-4}$). The dashed lines denote the Bayes adversarial error (left) and the Bayes standard error (right). Note that the adversarial error of estimators obtained from adversarial training, approaches the Bayes adversarial error as δ grows.

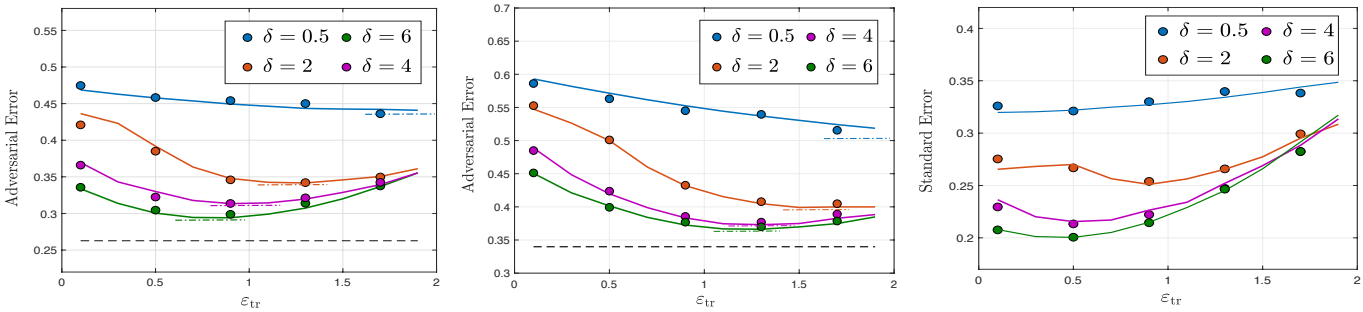

Fig. 2: Theoretical (solid lines) and Empirical (markers) results for the impact of adversarial training on the adversarial test error for $\varepsilon_{\rm ts} = 0.5$ (Left) and $\varepsilon_{\rm ts} = 0.9$ (Middle). The blacked dashed lines denote the Bayes adversarial error for the corresponding values of $\varepsilon_{\rm ts}$. The colored dashed lines depict the optimal value of each curve. Note that the optimal value of $\varepsilon_{\rm tr}$ decreases as δ grows. Right: Impact of adversarial training on the standard test error, illustrating that adversarial training can improve standard accuracy.

[HMRT19]), to the adversarial training case. Finally, we highlight the following observation from Figure 1 (right): For highly over-parametrized models (very small δ), standard accuracy remains the same for different choices of $\varepsilon_{\rm tr}$. As δ grows, adversarial training (perhaps surprisingly) seems to (also) improve the standard accuracy. However, for very large δ, increasing $\varepsilon_{\rm tr}$ hurts standard accuracy. These observations are consistent and theoretically validate corresponding findings on the role of data-set size on standard accuracy that were empirically observed in [TSE+18] for neural network training with non-synthetic datasets (e.g., MNIST).

b) Impact of $\boldsymbol{\varepsilon}_{\rm tr}$ on standard/adversarial errors.: Adversarial and Standard error curves based on the hyper-parameter $\varepsilon_{\rm tr}$ are illustrated in Figure 2. Note that the adversarial error behavior based on $\varepsilon_{\rm tr}$ is informative about the role of the data-set size on the optimal value of $\varepsilon_{\rm tr}$. The top figures show that the optimal value of $\varepsilon_{\rm tr}$ is typically larger than $\varepsilon_{\rm ts}$. Also note that as δ gets smaller, larger values of $\varepsilon_{\rm tr}$ are preferrable for robustness. Figure 2(Right) illustrates the impact of $\varepsilon_{\rm tr}$ on the standard error, where similar to Figure 1(Right), we observe that adversarial training can help standard accuracy. In particular, we observe that in the under-parameterized regime where $\delta > \delta \frac{\varepsilon_{\rm tr}}{\sqrt{n}}$ (as we will define in Section IV), adversarial training with small values of $\varepsilon_{\rm tr}$ is beneficial for accuracy. As δ increases, such gains diminish and indeed adversarial training starts hurting standard accuracy.

B. Proof Sketch

The complete proof of Theorem 1 is deferred to the long version of the paper [TPT20a]. Here, we provide an outline of the key steps in deriving (7) and (8).

a) Reducing (5) to a minimization problem.: For a decreasing loss function, picking $\boldsymbol{\delta}_i^\star \triangleq -y_i \, \mathtt{sign}(\boldsymbol{\theta}_n) \, \varepsilon_{\rm tr}/\sqrt{n}$, optimizes the inner maximization in (5). Therefore, (5) is equivalent to,

$$\min_{\boldsymbol{\theta}_n} \sum_{i=1}^{m} \mathcal{L}\Big(y_i \langle \boldsymbol{x}_i, \boldsymbol{\theta}_n \rangle - \frac{\varepsilon_{\rm tr}}{\sqrt{n}} \|\boldsymbol{\theta}_n\|_1\Big) + r \, \|\boldsymbol{\theta}_n\|_2^2. \qquad (9)$$

From (9), we can see now why the specific normalization of $\varepsilon_{\rm tr}$ is needed in (5). Recall that, $\boldsymbol{x}_i \sim \mathcal{N}(y_i \boldsymbol{\theta}_n^\star, \mathbb{I}_n)$ and $\|\boldsymbol{\theta}_n^\star\|_2 = 1$.

For fixed $\boldsymbol{\theta}$, the argument $y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle$ behaves as $\|\boldsymbol{\theta}\|_2(S+1)$, where $S \sim \mathcal{N}(0,1)$. Thus, for $\boldsymbol{\theta}$s that are such that $\|\boldsymbol{\theta}\|_2 = \Theta(1)$ (which ought to be the case for "good" classifiers in view of $\|\boldsymbol{\theta}_n^\star\|_2 = 1$), the term $y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle$ is an $\Theta(1)$-term. Now, thanks to the normalization $1/\sqrt{n}$ in (5), the second term $\frac{\varepsilon_{\rm tr}}{\sqrt{n}}\|\boldsymbol{\theta}\|_1$ in (9) is also of the same order. Here, we used again the intuition that $\|\boldsymbol{\theta}\|_1 = \Theta(\sqrt{n})$, as is the case for the true $\boldsymbol{\theta}^\star$. Our analysis formalizes these heuristic explanations.

b) The key statistics for the adversarial error: Our key observation is that the asymptotics of the adversarial error of a sequence of arbitrary classifiers $\{\boldsymbol{\theta}_n\}$, depend on the asymptotics of only a few key statistics of $\{\boldsymbol{\theta}_n\}$. This is formalized in the following lemma.

Lemma 2. *Define projection matrices Θ_n and $\Theta_n^\perp$ as $\Theta_n \triangleq \boldsymbol{\theta}_n^\star \boldsymbol{\theta}_n^{\star\top}/\|\boldsymbol{\theta}_n^\star\|_2^2$, $\Theta_n^\perp \triangleq \mathbb{I}_n - \Theta_n$. Assume that the sequence of $\{\boldsymbol{\theta}_n\}$ is such that the following limits hold*

$$\{\varepsilon_{tr}\,\|\boldsymbol{\theta}_n\|_1/\sqrt{n}\} \xrightarrow{P} w,\ \{\|\Theta_n\boldsymbol{\theta}_n\|_2\} \xrightarrow{P} \mu,\ \{\|\Theta_n^\perp\boldsymbol{\theta}_n\|_2\} \xrightarrow{P} \alpha,$$

Then, in the high-dimensional limit, the adversarial error satisfies

$$\mathcal{E}_{\varepsilon_{ts}}(\boldsymbol{\theta}_n) \xrightarrow{P} Q\Big(\frac{\mu - w\,\varepsilon_{ts}/\varepsilon_{tr}}{\sqrt{\mu^2+\alpha^2}}\Big) \tag{10}$$

Lemma 2 reduces the goal of computing asymptotics of the adversarial risk of $\widehat{\boldsymbol{\theta}}_n$ to computing asymptotics of the corresponding statistics $\|\widehat{\boldsymbol{\theta}}_n\|_1$, $\|\Theta_n\widehat{\boldsymbol{\theta}}_n\|_2$, and $\|\Theta_n^\perp\widehat{\boldsymbol{\theta}}_n\|_2$.

c) Scalarizing the objective function: The previous two steps set the stage for the core of the analysis, which we outline next. Thanks to step 1, we are now asked to analyze the statistical properties of a convex optimization problem. On top of that, due to step 2, the outcomes of the analysis ought to be asymptotic predictions for the quantities $\|\boldsymbol{\theta}_n\|_1$, $\|\Theta_n\boldsymbol{\theta}_n\|_2$ and $\|\Theta_n^\perp\boldsymbol{\theta}_n\|_2$. However, note that the term $\|\boldsymbol{\theta}_n\|_1$ appears inside the loss function. In particular, this is a new challenge, specific to robust optimization compared to previous analysis of standard regularized ERM. The first step to overcome these challenges is to identify an appropriate minimax Auxiliary Optimization (AO) problem that is probabilistically equivalent to (9). The second crucial step is to scalarize the AO based on an appropriate Lagrangian formulation. Finally, we perform a probabilistic analysis of the scalar AO. This results in the deterministic minimax problem in (7). See [TPT20a] for details.

IV. Robust Separability

An instance of special interest in practice is solving the *unregularized* version of the min-max problem:

$$\min_{\boldsymbol{\theta}_n}\ \frac{1}{m}\sum_{i=1}^{m}\max_{\|\boldsymbol{\delta}_i\|_\infty \le \varepsilon}\mathcal{L}\left(y_i\left\langle \boldsymbol{x}_i+\boldsymbol{\delta}_i, \boldsymbol{\theta}_n\right\rangle\right). \tag{11}$$

Following the same proof techniques as above, we can show that the formulas predicting the statistical behavior of this unconstrained version are given by the same formulas as in Theorem 1 with $r=0$ and also provided that the sampling ration δ is large enough so that a certain robust separability condition holds. In what follows, we describe this condition. We start with some background on (standard) data separability. Recall, that training data $\{(\boldsymbol{x}_i, y_i)\}$ are linearly separable if and only if $\exists \boldsymbol{\theta} \in \mathbb{R}^n$ such that for all training samples $y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle \ge 1$. Now, we say that data are $(\ell_\infty, \varepsilon)$-separable if and only if $\exists \boldsymbol{\theta} \in \mathbb{R}^n$ s.t. $y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle - \varepsilon\|\boldsymbol{\theta}\|_1 \ge 1,\ \forall i \in [m]$. Note that (standard) linear separability is equivalent to $(\ell_\infty, 0)$-separability as defined above. Moreover, it is clear that $(\ell_\infty, \varepsilon)$-separability implies $(\ell_\infty, 0)$-separability for any $\varepsilon \ge 0$. Recent works have shown that in the proportional limit data from stylized models are $(\ell_\infty, 0)$-separable if and only if the sampling ratio satisfies $\delta < \delta^*$ [CS18], [MRSY19] for some $\delta^* > 2$. We conjecture that there is a threshold δ_ε^*, depending on ε, such that data are $(\ell_\infty, \varepsilon)$-separable if and only if δ_ε^*. We believe that our techniques can be used to prove this conjecture and determine δ_ε^*, but we leave this interesting question to future work. Instead here, we simply note that based on the above discussion, if such a threshold exists, then it must satisfy $\delta_\varepsilon^* \le \delta^*$, for all values of ε, and in fact it is a decreasing function of ε. Now let us see how this notion relates to solving (5) and to our asymptotic characterization of its performance. Recall from (9) that the robust ERM for decreasing losses reduces to the minimization $\min_{\boldsymbol{\theta}} \sum_{i=1}^m \mathcal{L}(y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle - \varepsilon\|\boldsymbol{\theta}\|_1)$. Thus, using again the decreasing nature of the loss, it can be checked that the solution to the objective function above becomes unbounded for $\boldsymbol{\theta}$ such that the argument of the loss is positive for any $i \in [m]$. This is equivalent to the condition of $(\ell_\infty, \varepsilon)$-separability. In other words, when data are $(\ell_\infty, \varepsilon)$-separable, the robust estimator is unbounded. Recall from Section III-B that the minimax optimization variables w, μ, α represent the limits of $\|\boldsymbol{\theta}_n\|_1$, $\|\Theta_n\boldsymbol{\theta}_n\|_2$, and $\|\Theta_n^\perp\boldsymbol{\theta}_n\|_2$. Thus, if $\boldsymbol{\theta}_n$ is unbounded, then $w^\star, \mu^\star, \alpha^\star$ are not well defined. In accordance with this, we conjecture that the minimax problem (7) for $r=0$ (corresponding to (11)) has a solution if and only if the data are *not* $(\ell_\infty, \varepsilon)$-separable, equivalently, iff $\delta > \delta_\varepsilon^*$.

V. Conclusions and Future Directions

We studied the generalization behavior of adversarial training in a binary classification setting. In particular, we derived precise theoretical predictions for the performance of adversarial training for the Gaussian-mixture model. Numerical simulations validate theoretical predictions even for relatively small problem dimensions and demonstrate the role of all problem paramters on adversarial robustness. Finally, we remark that the current analysis can be extended to general convex regularization functions building on our ideas. An interesting future direction is analyzing adversarial training for Random Features and Neural Tangent Kernel models. One other natural question is considering attacks other than ℓ_q-norm attacks considered in this paper.

REFERENCES

[AZL20] Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. *arXiv preprint arXiv:2005.10190*, 2020.

[BCM19] Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems*, pages 7498–7510, 2019.

[BHMM18] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*, 2018.

[BHX19] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019.

[CAD+18] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

[CRS+19] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi. Unlabeled data improves adversarial robustness. *arXiv preprint arXiv:1905.13736*, 2019.

[CRWP19] Zachary Charles, Shashank Rajput, Stephen Wright, and Dimitris Papailiopoulos. Convergence and margin of adversarial training on separable data. *arXiv preprint arXiv:1905.09209*, 2019.

[CS18] Emmanuel J Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv preprint arXiv:1804.09753*, 2018.

[DHHR20] Edgar Dobriban, Hamed Hassani, David Hong, and Alexander Robey. Provable tradeoffs in adversarially robust classification. *arXiv preprint arXiv:2006.05161*, 2020.

[DL20] Oussama Dhifallah and Yue M Lu. A precise performance analysis of learning with random features. *arXiv preprint arXiv:2008.11904*, 2020.

[DL21] Oussama Dhifallah and Yue M Lu. On the inherent regularization effects of noise injection during training. *arXiv preprint arXiv:2102.07379*, 2021.

[DWR20] Chen Dan, Yuting Wei, and Pradeep Ravikumar. Sharp statistical guarantees for adversarially robust gaussian classification. *arXiv preprint arXiv:2006.16384*, 2020.

[GMKZ20] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020.

[GMMM19] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. In *NeurIPS*, 2019.

[GSS14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[HMRT19] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[JS20] Adel Javanmard and Mahdi Soltanolkotabi. Precise statistical analysis of classification accuracies for adversarial training. *arXiv preprint arXiv:2010.11213*, 2020.

[JSH20] Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. *arXiv preprint arXiv:2002.10477*, 2020.

[MDFF16] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[MM19] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[MMS+17] 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.

[MRSY19] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

[RW09] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[RXY+19] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

[RXY+20] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716*, 2020.

[SAH19] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. *arXiv preprint arXiv:1906.03761*, 2019.

[SN20] Samuel Henrique Silva and Peyman Najafirad. Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv preprint arXiv:2007.00753*, 2020.

[SST+18] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pages 5014–5026, 2018.

[Sto09] Mihailo Stojnic. Various thresholds for ℓ_1-optimization in compressed sensing. *arXiv preprint arXiv:0907.3666*, 2009.

[Sto13] Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

[SZS+13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arxiv 2013. *arXiv preprint arXiv:1312.6199*, 2013.

[TAH18] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[TOH15] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Proceedings of The 28th Conference on Learning Theory*, pages 1683–1709, 2015.

[TPT20a] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Asymptotic behavior of adversarial training in binary classification. *arXiv preprint arXiv:2010.13275*, 2020.

[TPT20b] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Sharp asymptotics and optimal performance for inference in binary models. In *International Conference on Artificial Intelligence and Statistics*, pages 3739–3749. PMLR, 2020.

[TPT21] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Fundamental limits of ridge-regularized empirical risk minimization in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 2773–2781. PMLR, 2021.

[TSE+18] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[ZYJ+19] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.