Asymptotic Behavior of Adversarial Training in Binary Linear Classification

Hossein Taheri, Ramtin Pedarsani, *Senior Member, IEEE*, and Christos Thrampoulidis

***Abstract*— Adversarial training using empirical risk minimization (ERM) 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 the generalization properties of adversarial training in classification remains widely open. In this article, 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 general ℓ_q-norm bounded perturbations ($q \geq 1$) in both discriminative binary models and generative Gaussian-mixture models with correlated features. 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 the fundamental limits of adversarial training.**

***Index Terms*— Adversarial learning, adversarial training, high-dimensional statistics, optimization.**

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 [31], [42]. 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., [11], [38] for a survey. Among many proposals in the already rich literature, perhaps the most popular approach is that of adversarial training [20]. 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 [25]. 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 article, we consider the adversarial training problem for ℓ_q-norm bounded perturbations in classification tasks, which solves the following robust empirical risk minimization (ERM) problem:

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

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 $\boldsymbol{\theta} \in \mathbb{R}^l$, ε_{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

$$\mathbb{E}_{\boldsymbol{x},y}\left[\max_{\|\boldsymbol{\delta}\|_q \leq \varepsilon_{\text{ts}}} \mathbf{1}_{\{y \neq \texttt{sign}(f_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x}+\boldsymbol{\delta}))\}}\right]$$

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, ε_{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 $\varepsilon_{\text{ts}} = 0$.

The goal of this article 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) [39], [40], [47] and in particular its applications to the convex ERM that enables its precise analysis, e.g., [30], [36], [43], [44], [46]. However, compared to previous works, we develop a new analysis for robust optimization with correlated data.

Manuscript received 12 July 2022; revised 17 April 2023; accepted 21 June 2023. This work was supported in part by the National Science Foundation under Grant 1909320, Grant 2003035, Grant 193464, and Grant 2009030; and in part by the GR8 award from King Abdullah University of Science and Technology. *(Corresponding author: Hossein Taheri.)*

Hossein Taheri and Ramtin Pedarsani are with the Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106 USA (e-mail: hossein@ucsb.edu; ramtin@ucsb.edu).

Christos Thrampoulidis is with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada, and also with the Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106 USA (e-mail: cthrampo@ece.ubc.ca).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TNNLS.2023.3290592.

Digital Object Identifier 10.1109/TNNLS.2023.3290592

Our main contributions are summarized as follows.

1) We precisely analyze, for the first time, the performance of adversarial training with ℓ_2 and ℓ_∞ attacks in binary classification for two important data models of Gaussian mixtures and generalized linear models. See Sections III and IV.
2) Our approach is general, allowing us to characterize the role of feature correlation, regularization, and general ℓ_q attacks with $q \geq 1$. In particular, our proof technique allows for nonisotropic features, yielding novel theoretical results even for nonadversarial convex regularized ERM settings (i.e., when $\varepsilon_{\text{tr}} = \varepsilon_{\text{ts}} = 0$). We elaborate on our technical approach in Section III-C.
3) Numerical illustrations in Section III-B 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 ε_{ts}, and the robust-optimization hyper-parameter ε_{tr} in our studied settings. Moreover, we observe interesting phenomena by comparing our results with the Bayes optimal robust errors.

A. Prior Works

Relevant to the flavor of our results, the recent work [24] studies precise tradeoffs and performance analysis in adversarial training with linear regression with ℓ_2 perturbations and isotropic Gaussian data. Compared to [24], our results hold for binary models, general ℓ_q perturbations with $q \geq 1$, and nonisotropic features with mild assumptions on the covariance matrix. 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 [23]. On the one hand, compared to [23] our analysis applies to both discriminative and generative data models, and also to the *regularized* ERM. Our analysis also allows generic covariance matrices while the analysis of [23] only applies to very specific structures for the covariance matrix. In addition, we examine how our formulas on adversarial training compare with those of the Bayes robust estimator. On the other hand, [23] extend their analysis to robust support vector machines (SVMs). Note, however, that we can retrieve the same results regarding the performance of adversarially-robust SVM by evaluating our formulas on regularized ERM with logistic loss and vanishing regularization parameters.

Our analysis of correlated features was motivated by [30], which derives sharp generalization guarantees for SVM models with correlated data. Very recently, correlated features have been considered in various settings, e.g., [8], [10], [15]. However, none of these works studies the more challenging problem of adversarially-robust ERM as we do here. To see, at a high level, why this 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 a robust error (rather than a standard error), and we show that this changes the statistics that need to be tracked by the CGMT analysis. Third, the primary optimization to which we eventually apply the CGMT involves an "effective" ℓ_p-regularizer (where ℓ_p is the dual norm of the adversary's ℓ_q-norm), which unlike previous works appears inside the argument of the loss function, requiring new techniques to scalarize the auxiliary optimization (AO). Specially, we do this in the presence of nonisotropic features, which yields new results even for standard ERM methods.

The adversarial Bayes risk for Gaussian mixtures has been recently characterized in [5], [13], and [17]. Here, we combine their results with our precise asymptotics on the practically relevant adversarial training method, allowing us to investigate the fundamental limits of the latter. Allen-Zhu and Li [12] and Charles et al. [1] discuss the optimization landscape of adversarial training, however, these works do not address the generalization properties of adversarial training, as done in this article. The prior work [29] considers adversarial training with linear loss to analyze the sample complexity of robust estimators. Instead, here, we investigate the more challenging, but practically more relevant, 0/1-loss and its tractable approximations (e.g., hinge and logistic). Another related line of work studies tradeoffs between the standard and adversarial errors, e.g., [17], [33], [48], [49], but for simpler algorithms and data models, rather than adversarial training and correlated GLM/GMM, which we focus on here. The benefits of unlabeled data in robustness have been investigated in several works, e.g., [7], [32]. 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 [15], [16], [18], [19], [28]. A preliminary version of this work appeared in [45]. The results presented in [45] only apply to data that follow the isotropic Gaussian mixture model and only to ℓ_∞ attacks. The current manuscript significantly extends the scope of these results: first, we extend the results for GMM to general covariance matrices (not necessarily isotropic). This is important because it better captures data distributions in practice. We also note that the extension is technically nontrivial, requiring several modifications in the proof compared with the isotropic case. Second, in the journal version, we describe a unifying analysis and results that apply both to discriminative and generative models. Specifically for discriminative models, we present new results for GLM data. Third, we provide a general analysis of ℓ_p-norm attacks. This extends the results of the conference version that only applied to ℓ_∞-norm. For demonstration, we present results for ℓ_2-attacks in Section IV. Finally, we have extended our numerical study by introducing additional experiments in Appendix VII, as shown in the supplementary material.

Notation: Letting $\delta(x)$ denote a Dirac delta mass at x, the empirical distribution of a vector $\boldsymbol{x} \in \mathbb{R}^n$ is given by $(1/n)\sum_{i=1}^n \delta(\boldsymbol{x}_i)$. The empirical joint distribution of $\mathbf{v}, \mathbf{u} \in \mathbb{R}^n$ is given by $(1/n)\sum_{i=1}^n \delta(\mathbf{v}_i, \mathbf{u}_i)$. The Wasserstein-$k$ distance

between two measures ρ_1, ρ_2 is defined as $W_k(\rho_1, \rho_2) \triangleq (\inf_{\rho\in P} \mathbb{E}_{(X,Y)\sim\rho}|X-Y|^k)^{1/k}$, where P denotes all couplings of ρ_1 and ρ_2. We say that a sequence of probability distributions μ_n converges in Wasserstein-k distance to a probability distribution μ, if $W_k(\mu_n, \mu) \to 0$ as $n \to \infty$. The Gaussian Q-function is denoted by $Q(\cdot)$. $\odot$ denotes the elementwise multiplication. The function $\|\cdot\|_q^p$ is denoted by ℓ_q^p. For a positive semidefinite matrix S, we define $\|\mathbf{v}\|_S \triangleq (\mathbf{v}^T S\mathbf{v})^{1/2}$. Finally, for a sequence of random variables $X_{m,n}$ that converges in probability to some constant c in the proportional asymptotic limit $m, n \to \infty$, $m/n \to \delta$, we write $X_{m,n} \xrightarrow{P} c$.

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 two stylized models for binary classification.

1) Gaussian Mixture Models: The first model is a 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 $y_i \in \{\pm 1\}$) and with covariance $\boldsymbol{\Sigma}_n$. The subscript n emphasizes the dependence on dimension. Formally, the GMM assumes

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

2) Generalized Linear Models: The second model is a generalized linear model (GLM) with a binary link function. Specifically, assume that the label $y_i \in \{\pm 1\}$ associated with the feature vector $\boldsymbol{x}_i$ is generated as

$$y_i = \psi\big(\langle\boldsymbol{\theta}_n^\star, \boldsymbol{x}_i\rangle\big), \quad \boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_n) \tag{3}$$

for a possibly random link function $\psi{:}\mathbb{R} \to \{\pm 1\}$. This includes the well-known logistic and signed models, by letting $\mathbb{P}(\psi(x) = 1) = 1/(1+\exp(-x))$ and $\psi(x) = \mathtt{sign}(x)$, respectively.

We assume that the underlying (unknown) vector of regressors $\boldsymbol{\theta}_n^\star \in \mathbb{R}^n$, and the covariance matrix $\boldsymbol{\Sigma}_n \in \mathbb{R}^{n\times n}$, satisfy the following technical (and mild) assumptions.

Assumption 1: The minimum and maximum eigenvalues of the covariance matrices $\boldsymbol{\Sigma}_n$ satisfy $0 < c < \lambda_{\min}(\boldsymbol{\Sigma}_n)$ and $\lambda_{\max}(\boldsymbol{\Sigma}_n) < C < \infty$.

Assumption 2: Denoting $\zeta_n \triangleq (\boldsymbol{\theta}_n^{\star\top}\boldsymbol{\Sigma}_n\boldsymbol{\theta}_n^\star)^{1/2}$ for GLM and $\widetilde{\zeta}_n \triangleq (\boldsymbol{\theta}_n^{\star\top}\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\theta}_n^\star)^{1/2}$ for GMM, we define their high-dimensional limits as ζ and $\widetilde{\zeta}$, i.e., $\zeta_n \xrightarrow{P} \zeta$ and $\widetilde{\zeta}_n \xrightarrow{P} \widetilde{\zeta}$. Moreover, for both models, we assume without loss of generality that $\|\boldsymbol{\theta}_n^\star\|_2 \xrightarrow{P} 1$.

Assumption 3: Let $\boldsymbol{\Sigma}_n = \mathbf{U}_n\Lambda_n\mathbf{U}_n^\top$ be the eigendecomposition of $\boldsymbol{\Sigma}_n$ and let $\lambda_{n,i}$ denote the ith entry on the diagonal of $\Lambda_\mathbf{n}$. Denote $\mathbf{v}_n \triangleq \mathbf{U}_n^\top\boldsymbol{\theta}_n^\star$. Then, the joint distribution of $(\sqrt{n}\boldsymbol{\theta}_{n,i}^\star, \lambda_{n,i}, \sqrt{n}\mathbf{v}_{n,i})$, $i \in [n]$, converges in Wasserstein-2 distance to a probability distribution Π in $\mathbb{R}\times\mathbb{R}_+\times\mathbb{R}$, i.e.,

$$\frac{1}{n}\sum_{i=1}^n \delta\big(\sqrt{n}\boldsymbol{\theta}_{n,i}^\star, \lambda_{n,i}, \sqrt{n}\mathbf{v}_{n,i}\big) \xrightarrow{W_2} \Pi.$$

The assumption on $\|\boldsymbol{\theta}_n^\star\|_2$ is without loss of generality for GLM since $\|\boldsymbol{\theta}_n^\star\|_2$ can be absorbed in the link function ψ. Similarly for GMM, if $\|\boldsymbol{\theta}_n^\star\|_2 \neq 1$, we can always assume normalized features $\boldsymbol{x}$, by appropriately scaling the covariance matrix $\boldsymbol{\Sigma}_n$. We remark that while the Gaussian distribution assumption on feature vectors is crucial for our theoretical analysis, our empirical results suggest that this assumption can be relaxed to include at least the family of sub-Gaussian data distributions. We discuss this universality property in Appendix VII-B (supplementary material).

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 either model (2) or (3). As is typical, given $\widehat{\boldsymbol{\theta}}_n$, a decision is made about the label of $\boldsymbol{x}$ based on $\mathtt{sign}(\langle\boldsymbol{x}, \widehat{\boldsymbol{\theta}}_n\rangle)$. Thus, letting y be the label of a fresh sample $\boldsymbol{x}$, the *standard error* is given by

$$\mathcal{E}\big(\widehat{\boldsymbol{\theta}}_n\big) \triangleq \mathbb{E}_{\boldsymbol{x},y}\Big[\mathbf{1}_{\{y\neq\mathtt{sign}(\langle\boldsymbol{x},\widehat{\boldsymbol{\theta}}_n\rangle)\}}\Big]. \tag{4}$$

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

$$\mathcal{E}_{\ell_q,\varepsilon_{\text{ts}}}(\widehat{\boldsymbol{\theta}}_n) \triangleq \mathbb{E}_{\boldsymbol{x},y}\Big[\max_{\|\boldsymbol{\delta}\|_q\leq\varepsilon_{\text{ts}}} \mathbf{1}_{\{y\neq\mathtt{sign}(\langle\boldsymbol{x}+\boldsymbol{\delta},\widehat{\boldsymbol{\theta}}_n\rangle)\}}\Big]. \tag{5}$$

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 this article, we assume that $\mathcal{L}$ is convex and decreasing. This includes popular choices such as the logistic, hinge, and exponential losses.

III. Main Results for ℓ_∞ Perturbations

A. Asymptotic Behavior

In this section, we focus on the case of bounded ℓ_∞-perturbations, i.e., the adversarial error in (5) is considered for $q = \infty$. Specifically, let $\widehat{\boldsymbol{\theta}}_n$ be a solution to the following robust minimization:

$$\min_{\boldsymbol{\theta}_n}\sum_{i=1}^m \max_{\|\boldsymbol{\delta}_i\|_\infty\leq\frac{\varepsilon_{\text{tr}}}{\sqrt{n}}} \mathcal{L}(y_i\langle\boldsymbol{x}_i+\boldsymbol{\delta}_i, \boldsymbol{\theta}_n\rangle) + r\|\boldsymbol{\theta}_n\|_2^2. \tag{6}$$

In our asymptotic setting, ε_{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 $\|\boldsymbol{\delta}_i\|_2$, is comparable to the norm of the true vector $\|\boldsymbol{\theta}_n^\star\|_2$, i.e., both are constant in the high-dimensional limit $\to n$. We explain this normalization

further in Section III-C. Here, we consider the case of the diagonal covariance matrix (i.e., $\boldsymbol{\Sigma}_n = \Lambda_n$). Note that this assumption can be made without loss of generality for GMM data. Indeed, instead of features $\boldsymbol{x}_i$ as in (2) and mean vector $\boldsymbol{\theta}_n^\star$, we can equivalently analyze features $\widetilde{\boldsymbol{x}}_i = \mathbf{U}_n^\top \boldsymbol{x}_i$ and mean vector $\widetilde{\boldsymbol{\theta}}_n^\star := \mathbf{U}_n^\top \boldsymbol{\theta}_n^\star$. For the GLM data in (3), we defer the general case of possibly nondiagonal $\boldsymbol{\Sigma}_n$ to Appendix VIII, as shown in the supplementary material, where we also discuss how final expressions simplify in the case of isotropic features.

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

$$\mathcal{M}_f(x;\kappa) \triangleq \min_v \frac{1}{2\kappa}(x-v)^2 + f(v) \tag{7}$$

for the Moreau envelope of a function $f{:}\mathbb{R} \to \mathbb{R}$ at $x \in \mathbb{R}$ with parameter $\kappa > 0$ [35]. 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:

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

where $C = \widetilde{\zeta}$ and ζ (defined in Assumption 2) for GMM and GLM, respectively. We introduce the following min–max objective based on the eight scalars:

$$\min_{\substack{\alpha,\tau_1,w\in\mathbb{R}_+,\\ \mu\in\mathbb{R}}} \max_{\substack{\tau_2,\beta,\gamma\in\mathbb{R}_+,\\ \eta\in\mathbb{R}}} f(\bar{\mathbf{v}}) + \mathbb{E}\left[\mathcal{M}_{\mathcal{L}}\left(Z_{\alpha,\mu} - w; \frac{\tau_1}{\beta}\right)\right] + \varepsilon_{\mathrm{tr}}\gamma\, \mathbb{E}\left[\mathcal{M}_{\ell_1+\frac{r}{\varepsilon_{\mathrm{tr}}\gamma}\ell_2^2}\left(\frac{\alpha\beta}{\tau_2\sqrt{\delta L}}H + \frac{\alpha\eta}{\tau_2 D}\,T; \frac{\alpha\varepsilon_{\mathrm{tr}}\gamma}{\tau_2 L}\right)\right] \tag{8}$$

where $D \triangleq \widetilde{\zeta}^2 L$ and ζ^2 for models (2) and (3), respectively, $H \sim \mathcal{N}(0,1)$ and $(T, L, V) \sim \Pi$ where Π was defined in Assumption 3. We also let for convenience

$$Z_{\alpha,\mu} \triangleq \begin{cases} \sqrt{\alpha^2 + \mu^2\widetilde{\zeta}^2}\, G + \mu\widetilde{\zeta}^2, & \text{for GMM} \\ \alpha G + \mu\zeta S \cdot \psi(\zeta S), & \text{for GLM} \end{cases} \tag{9}$$

where $G, S \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Notice that the objective function of (8) depends explicitly on the sampling ratio δ and on the training parameter $\varepsilon_{\mathrm{tr}}$. Moreover, it depends implicitly on $\boldsymbol{\theta}_n^\star$ and Λ_n via T and L, respectively, and on the specific loss $\mathcal{L}$ via its Moreau envelope. The nature of allowed perturbations (the ℓ_∞-type) is reflected in (8), 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 (8) and the adversarial risk of the robust classifier $\widehat{\boldsymbol{\theta}}_n$. The proof is deferred to Appendices VIII-C and VIII-B (supplementary material).

Theorem 1: Assume that the training dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, is generated according to either (2) or (3) with diagonal covariance matrices satisfying Assumptions 1–3. Consider the robust classifiers $\{\widehat{\boldsymbol{\theta}}_n\}$, obtained by adversarial training in (6). Then, the high-dimensional limit for the adversarial error $\mathcal{E}_{\ell_\infty,(\varepsilon_{\mathrm{ts}}/\sqrt{n})}(\widehat{\boldsymbol{\theta}}_n)$, converges to

$$\begin{cases} Q\left(\dfrac{\mu^\star\widetilde{\zeta}^2 - w^\star\,\varepsilon_{\mathrm{ts}}/\varepsilon_{\mathrm{tr}}}{\sqrt{\mu^{\star 2}\widetilde{\zeta}^2 + \alpha^{\star 2}}}\right), & \text{for GMM} \\ \mathbb{P}\left(\mu^\star\zeta\, S\,\psi(\zeta S) + \alpha^\star G < \dfrac{w^\star\varepsilon_{\mathrm{ts}}}{\varepsilon_{\mathrm{tr}}}\right), & \text{for GLM} \end{cases} \tag{10}$$

where $Q(\cdot)$ denotes the Gaussian Q-function and $(\alpha^\star, \mu^\star, w^\star)$ is the unique solution to the scalar min–max problem (8).

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 Section III-B, we demonstrate the precise theoretical values and the corresponding numerical values.

B. 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 $\varepsilon_{\mathrm{tr}}$ and $\varepsilon_{\mathrm{ts}}$. 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; thus L has a unit mass at 1. Additional experiments on GLM are given in Appendix VII-A, as shown in the supplementary material. We further assume that T is standard normal and fix regularization parameter $r = 10^{-4}$. To solve (8), we derive the solution of the corresponding saddle point equations (derived in (60) in Appendix VIII-D1, as shown in the supplementary material) 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 (6) by gradient descent. The resulting estimator is used to compute the adversarial test error by evaluating (4) 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 shown in Figs. 1 and 2. Next, we discuss some of the insights obtained from these figures.

1) Impact of $\boldsymbol{\delta}$ on Standard/Adversarial Errors: Fig. 1 shows the adversarial and standard errors as a function of $\delta = m/n$. We compare the results of adversarial training with the Bayes optimal error. Formally, the Bayes adversarial error is defined as

$$\mathcal{E}_{\ell_q,\varepsilon_{\mathrm{ts}}}(\mathrm{OPT}) \triangleq \min_{f_{\boldsymbol{\theta}}} \mathbb{E}_{\boldsymbol{x},y}\left[\max_{\|\boldsymbol{\delta}\|_q \le \varepsilon_{\mathrm{ts}}} \mathbf{1}_{\{y \neq f_{\boldsymbol{\theta}}(\boldsymbol{x}+\boldsymbol{\delta})\}}\right]. \tag{11}$$

For the Gaussian-mixture model (2) under an ℓ_q attack with budget ε, the Bayes adversarial error is derived as follows [5]:

$$\mathcal{E}_{\ell_q,\varepsilon}(\mathrm{OPT}) = Q\left(\|\boldsymbol{\theta}^\star - \boldsymbol{\mu}^\star\|_{\boldsymbol{\Sigma}_n^{-1}}\right), \quad \text{where } \boldsymbol{\mu}^\star \triangleq \arg\min_{\|\boldsymbol{\mu}\|_q \le \varepsilon} \|\boldsymbol{\theta}^\star - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}_n^{-1}}. \tag{12}$$

The dashed lines in Fig. 1 show the *Bayes adversarial error*, derived according to (12) for $\varepsilon = \varepsilon_{\mathrm{ts}}/\sqrt{n}$.

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 $\varepsilon_{\mathrm{ts}}$. In Appendix X-C, as shown in the supplementary material, we formally prove that for ℓ_2 attacks bounded by $\varepsilon_{\mathrm{ts}} \in [0, 1]$, the robust error achieved by adversarial training with any $\varepsilon_{\mathrm{tr}} \in [0, 1]$

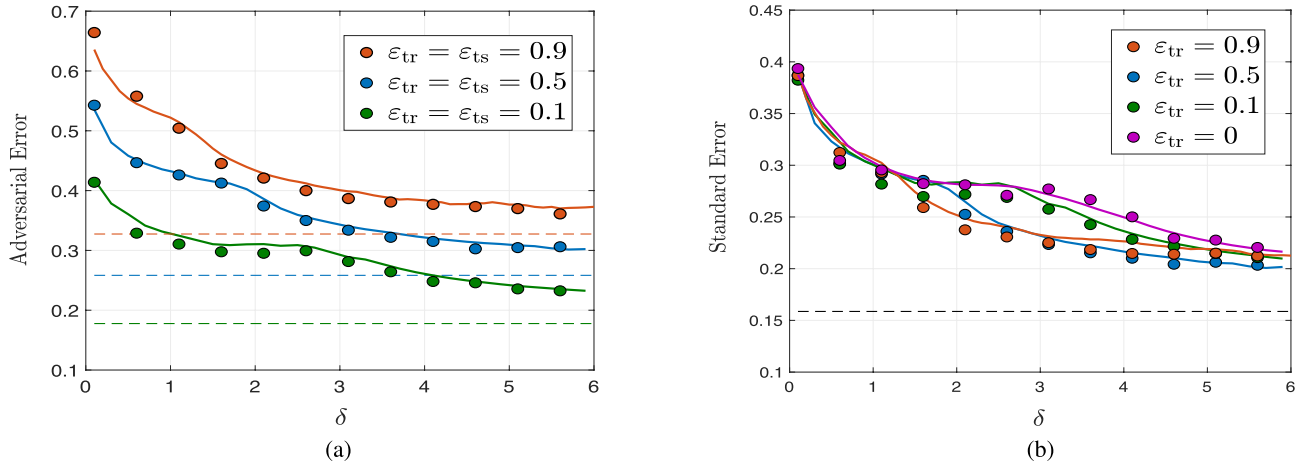

Fig. 1. Adversarial/standard test error based on $\delta := m/n$. Solid lines correspond to theoretical predictions while markers denote the empirical results derived by solving ERM with vanishing regularization ($r = 10^{-4}$) using gradient descent. The dashed lines denote (a) Bayes's adversarial error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

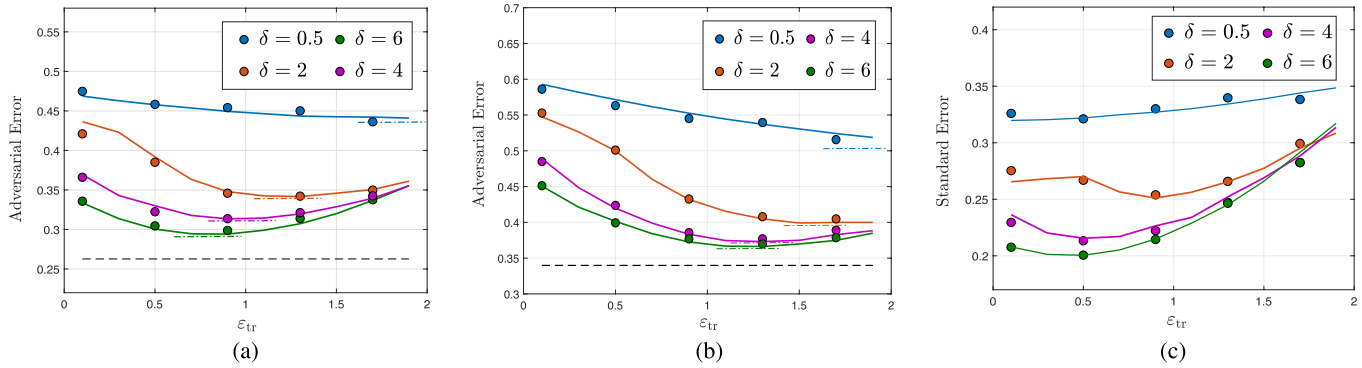

Fig. 2. Theoretical (solid lines) and empirical (markers) results for the impact of adversarial training on the adversarial test error for (a) $\varepsilon_{\rm ts} = 0.5$ and (b) $\varepsilon_{\rm ts} = 0.9$. The black dashed lines denote the Bayes adversarial error for the corresponding values of $\varepsilon_{\rm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

converges to the Bayes adversarial error in the infinite sample-size limit, i.e., when $\delta \to \infty$. In general, in light of the comparison between the error formulas of Theorem 1 and the Bayes adversarial error, Fig. 1 provides a means to quantify the suboptimality 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 [37], 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 a separability threshold $\delta_{(\varepsilon_{\rm tr}/\sqrt{n}),\Pi}$, which we define as the maximum value of δ for which the data-points are $(\ell_\infty, (\varepsilon_{\rm tr}/\sqrt{n}))$-separable (for definition, see the discussion on robust separability in Section V). 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 [24], 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 an extension of the more familiar double-descent behavior in standard ERM (first observed in numerous high-dimensional machine learning models [3], [4], [21]), to the adversarial training case. Finally, we highlight the following observation from Fig. 1(a): for highly over-parameterized 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 dataset size on standard accuracy that was empirically observed in [48] for neural network training with nonsynthetic datasets such as MNIST.

2) 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 Fig. 2. Note that the adversarial error behavior based on $\varepsilon_{\rm tr}$ is informative about the role of the dataset size on the optimal value of $\varepsilon_{\rm tr}$. Fig. 2(a) and (b) shows 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 preferable for robustness. As detailed in Appendix VII-C, as shown in the supplementary material, this behavior is also observed in real-world experiments with the MNIST dataset. Fig. 2(c) illustrates the impact of $\varepsilon_{\rm tr}$ on the standard error,

where similar to Fig. 1(b), we observe that adversarial training can help standard accuracy. In particular, we observe that in the under-parameterized regime where $\delta > \delta_{(\varepsilon_{\rm tr}/\sqrt{n}),\Pi}$ (as we will define in Section V), 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.

C. Proof Sketch

The complete proof of Theorem 1 is deferred to the appendix (supplementary material). Here, we provide an outline of the key steps in deriving (8) and (10).

1) Reducing (6) 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 (6). Therefore, (6) is equivalent to

$$\min_{\boldsymbol{\theta}_n} \sum_{i=1}^{m} \mathcal{L}\Big(y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}_n\rangle - \frac{\varepsilon_{\rm tr}}{\sqrt{n}}\|\boldsymbol{\theta}_n\|_1\Big) + r\|\boldsymbol{\theta}_n\|_2^2. \tag{13}$$

From (13), we can see now why the specific normalization of $\varepsilon_{\rm tr}$ is needed in (6). Recall that (for model (3), for instance), $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_n)$ and $\|\boldsymbol{\theta}_n^\star\|_2 \xrightarrow{P} 1$. For simplicity assume here that $\boldsymbol{\Sigma}_n = \mathbb{I}_n$. For fixed $\boldsymbol{\theta}$, the argument $y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle$ behaves as $\|\boldsymbol{\theta}\|_2 S\psi(S)$, 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 (6), the second term $(\varepsilon_{\rm tr}/\sqrt{n})\|\boldsymbol{\theta}\|_1$ in (13) 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.

2) 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. The proof is deferred to Appendix VIII-A (supplementary material). Similar to before, there is nothing special here to $q = \infty$, so we state this result for general q.

Lemma 1: Fix $q \geq 1$ and let ℓ_p denote the dual norm of ℓ_q. Let $\widetilde{\boldsymbol{\theta}}_n^\star \triangleq \boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\theta}_n^\star$ for data model (3) and $\widetilde{\boldsymbol{\theta}}_n^\star \triangleq \boldsymbol{\Sigma}_n^{-1/2}\boldsymbol{\theta}_n^\star$ for data model (2). Furthermore, for both models, define projection matrices Θ_n and $\Theta_n^\perp$ as follows, $\Theta_n \triangleq \widetilde{\boldsymbol{\theta}}_n^\star\widetilde{\boldsymbol{\theta}}_n^{\star\top}/\|\widetilde{\boldsymbol{\theta}}_n^\star\|_2^2$, $\Theta_n^\perp \triangleq \mathbb{I}_n - \Theta_n$. Furthermore, let ε and ε' (possibly scaling with the problem dimensions) be the upper bounds on the norm of the adversarial perturbation during training and test time, respectively. With this notation, assume that the sequence of $\{\boldsymbol{\theta}_n\}$ is such that the following limits are true for the statistics $\|\boldsymbol{\theta}_n\|_p$, $\|\Theta_n\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\theta}_n\|_2$ and $\|\Theta_n^\perp\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\theta}_n\|_2$:

$$\big\{\varepsilon\|\boldsymbol{\theta}_n\|_p\big\} \xrightarrow{P} w, \quad \frac{1}{C}\big\{\big\|\Theta_n\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\theta}_n\big\|_2\big\} \xrightarrow{P} \mu$$
$$\big\{\big\|\Theta_n^\perp\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\theta}_n\big\|_2\big\} \xrightarrow{P} \alpha$$

where $C = \tilde{\zeta}, \zeta$ for GMM and GLM, respectively. Then, in the high-dimensional limit, the adversarial error converges to

$$\begin{cases} Q\Big(\dfrac{\mu\widetilde{\zeta}^2 - w\,\varepsilon'/\varepsilon}{\sqrt{\mu^2\widetilde{\zeta}^2 + \alpha^2}}\Big), & \text{for GMM} \\ \mathbb{P}\big(\mu\zeta S\,\psi(\zeta S) + \alpha G - w\varepsilon'/\varepsilon < 0\big), & \text{for GLM.} \end{cases} \tag{14}$$

The detailed proof of the lemma is deferred to the appendix (supplementary material). There are essentially two steps in establishing the result. The first is to exploit the decreasing nature of the 0/1-loss to explicitly optimize over $\boldsymbol{\delta}_i$. This optimization gives rise to the dual norm $\|\boldsymbol{\theta}_n\|_p$. The second step is to consider the change of variables $\boldsymbol{\theta}_n \Rightarrow \widetilde{\boldsymbol{\theta}}_n \triangleq \boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\theta}_n$ and decompose $\widetilde{\boldsymbol{\theta}}_n$ on its projection on $\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\theta}_n^\star$ and its complement. In the notation of the lemma, $\widetilde{\boldsymbol{\theta}}_n = \Theta_n\widetilde{\boldsymbol{\theta}}_n + \Theta_n^\perp\widetilde{\boldsymbol{\theta}}_n$. The Gaussianity of the feature vectors together with orthogonality of the two components in the decomposition of $\boldsymbol{\theta}_n$ explain the appearance of the Gaussian variables S and G in (14). When applied to ℓ_∞-perturbations, Lemma 1 reduces the goal of computing asymptotics of the adversarial risk of $\widehat{\boldsymbol{\theta}}_n$ to computing asymptotics of the corresponding statistics $\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}}_n\|_1$, $\|\Theta_n\widetilde{\boldsymbol{\theta}}_n\|_2$, and $\|\Theta_n^\perp\widetilde{\boldsymbol{\theta}}_n\|_2$.

3) 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 $\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}}_n\|_1$, $\|\Theta_n\widetilde{\boldsymbol{\theta}}_n\|_2$, and $\|\Theta_n^\perp\widetilde{\boldsymbol{\theta}}_n\|_2$. However, note that the term $\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}}_n\|_1$ appears inside the loss function. In particular, this is a new challenge, specific to robust optimization compared with previous analyses of standard regularized ERM. Moreover, both of the terms $\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}}_n\|_1$ and $\|\boldsymbol{\Sigma}_n^{-1/2}\widetilde{\boldsymbol{\theta}}_n\|_2^2$ are not decomposable based on $\|\Theta_n\widetilde{\boldsymbol{\theta}}_n\|_2$ and $\|\Theta_n^\perp\widetilde{\boldsymbol{\theta}}_n\|_2$, due to the presence of the term $\boldsymbol{\Sigma}_n^{-1/2}$. The first step to overcome these challenges is to identify an appropriate min–max AO problem that is probabilistically equivalent to (13). 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 min–max problem in (8). See the appendix (supplementary material) for details.

IV. Main Results for ℓ_2 Perturbations

When $q = 2$, the min-max problem is equivalent to the following, by choosing the optimal choice $\boldsymbol{\delta}_i = -y_i\varepsilon_{\rm tr}\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2$:

$$\min_{\boldsymbol{\theta}_n} \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}_n\rangle - \varepsilon_{\rm tr}\|\boldsymbol{\theta}_n\|_2) + r\|\boldsymbol{\theta}_n\|_2^2. \tag{15}$$

Here, we assume $\{\boldsymbol{\Sigma}_n\}$ to be a sequence of positive definite matrices. Denote $\widetilde{\mathbf{v}} \triangleq (\alpha, \tau_1, \tau_3, w, \mu, \tau_2, \beta, \gamma, \eta)$ and define $g{:}\mathbb{R}^9 \to \mathbb{R}$ as follows:

$$g(\widetilde{\mathbf{v}}) \triangleq -\gamma w - \frac{\mu^2\tau_2}{2\alpha}C^2 - \frac{\alpha\beta^2}{2\delta\tau_2} - \frac{\alpha\tau_2}{2} + \frac{\beta\tau_1}{2} + \eta\mu - \frac{\eta^2\alpha}{2\tau_2 C^2} + \frac{\varepsilon_{\rm tr}\gamma\tau_3}{2}$$

where recall that $C \triangleq \widetilde{\zeta}$ and ζ for GMM and GLM, respectively. With this notation, we introduce the following

min–max problem:

$$\min_{\substack{\alpha,\tau_1,\tau_3,w\in\mathbb{R}_+\\ \mu\in\mathbb{R}}} \max_{\substack{\tau_2,\beta,\gamma\in\mathbb{R}_+,\\ \eta\in\mathbb{R}}} g(\widetilde{\mathbf{v}}) + \mathbb{E}\left[\mathcal{M}_{\mathcal{L}}\left(Z_{\alpha,\mu} - w; \frac{\tau_1}{\beta}\right)\right] + \frac{\eta^2\alpha^2}{\tau_2^2 C^4}\left(\frac{\varepsilon_{\rm tr}\gamma}{2\tau_3} + r\right)\mathbb{E}_L\left[\frac{\frac{C^4\beta^2}{\eta^2\delta} + \widetilde{L}}{\frac{\varepsilon_{\rm tr}\gamma\alpha + 2\tau_3 r\alpha}{\tau_2\tau_3} + L}\right] \tag{16}$$

where we define $\widetilde{L} \triangleq 1/L$ and L for GMM and GLM, respectively, and the random variables L and $Z_{\alpha,\mu}$ are defined the same as in (8).

Theorem 2: Consider the same setting as in Theorem 1, only here assume that $q = 2$ and $\{\boldsymbol{\Sigma}_n\}$ are positive definite matrices (not necessarily diagonal) satisfying Assumptions 1–3. Let $(\alpha^\star, \mu^\star, w^\star)$ be the unique solution to the min–max problem (16). Then, the high-dimensional limit for the adversarial error $(\mathcal{E}_{\ell_2,\varepsilon_{\rm ts}}(\widehat{\boldsymbol{\theta}}_n))$ converges to

$$\begin{cases} Q\left(\dfrac{\mu^\star\widetilde{\zeta}^2 - w^\star\,\varepsilon_{\rm ts}/\varepsilon_{\rm tr}}{\sqrt{\mu^{\star 2}\widetilde{\zeta}^2 + \alpha^{\star 2}}}\right), & \text{for GMM} \\ \mathbb{P}\left(\mu^\star\zeta S\psi(\zeta S) + \alpha^\star G < w^\star\dfrac{\varepsilon_{\rm ts}}{\varepsilon_{\rm tr}}\right), & \text{for GLM.} \end{cases} \tag{17}$$

Proof of Theorem 2 is deferred to Appendix IX-A, as shown in the supplementary material. Compared to Theorem 1, note here that the asymptotic prediction only depends on the total energy of $\boldsymbol{\theta}_n^\star$ (which was assumed to be 1 in Assumption 2) and not on its empirical distribution T. We present numerical illustrations on ℓ_2-attacks in Appendix IX-A, as shown in the supplementary material, where we also discuss how the dataset size and attack budgets, affect the adversarial and standard test errors based on Theorem 2.

V. Further Discussion on Our Results

Remark 1 (Training With No Regularization and Robust Separability*):* An instance of special interest in practice is solving the *unregularized* version of the min–max problem

$$\min_{\boldsymbol{\theta}_n} \frac{1}{m}\sum_{i=1}^{m} \max_{\|\boldsymbol{\delta}_i\|_q \le \varepsilon} \mathcal{L}(y_i\langle \boldsymbol{x}_i + \boldsymbol{\delta}_i, \boldsymbol{\theta}_n\rangle). \tag{18}$$

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 (ℓ_q, ε)-separable if and only if

$$\exists\boldsymbol{\theta} \in \mathbb{R}^n, \quad \text{s.t. } y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle - \varepsilon\|\boldsymbol{\theta}\|_p \ge 1 \quad \forall i \in [m].$$

Note that (standard) linear separability is equivalent to $(\ell_q, 0)$-separability as defined above. Moreover, it is clear that (ℓ_q, ε)-separability implies $(\ell_q, 0)$-separability for any $\varepsilon \ge 0$. Recent works have shown that in the proportional limit data from the GLM are $(\ell_q, 0)$-separable if and only if the sampling ratio satisfies $\delta < \delta_\psi$ [6], [14], [30], [41] for some $\delta_\psi > 2$. Here, the subscript ψ denotes dependence of the phase-transition threshold δ_ψ on the link function ψ of the GLM. We conjecture that there is a threshold $\delta_{\psi,\varepsilon,\Pi}$, depending on ε, the link function ψ and the probability distribution Π such that data are (ℓ_q, ε)-separable if and only if $\delta < \delta_{\psi,\varepsilon,\Pi}$. We believe that our techniques can be used to prove this conjecture and determine $\delta_{\psi,\varepsilon,\Pi}$, 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 $\delta_{\psi,\varepsilon,\Pi} \le \delta_{\psi,0,\Pi}$, 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 (6) and to our asymptotic characterization of its performance. Recall from (13) that the robust ERM for decreasing losses reduces the minimization $\min_{\boldsymbol{\theta}} \sum_{i=1}^m \mathcal{L}(y_i\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle - \varepsilon\|\boldsymbol{\theta}\|_p)$. 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 (ℓ_q, ε)-separability. In other words, when data are (ℓ_q, ε)-separable, the robust estimator is unbounded. Recall from Section III-C that the min–max optimization variables w, μ, α represent the limits of $\|\widehat{\boldsymbol{\theta}}_n\|_p$, $\|\Theta_n\boldsymbol{\Sigma}_n^{1/2}\widehat{\boldsymbol{\theta}}_n\|_2$, and $\|\Theta_n^\perp\boldsymbol{\Sigma}_n^{1/2}\widehat{\boldsymbol{\theta}}_n\|_2$. Thus, if $\widehat{\boldsymbol{\theta}}_n$ is unbounded, then $w^\star, \mu^\star, \alpha^\star$ are not well defined. In accordance with this, we conjecture that the min–max problem (8) for $r = 0$ [corresponding to (18)] has a solution if and only if the data are *not* (ℓ_q, ε)-separable, equivalently, if $\delta > \delta_{\psi,\varepsilon,\Pi}$. Equivalent results are applicable to the Gaussian-Mixture models.

Remark 2 (On Statistical Limits in Adversarial Training): The asymptotics in (17) imply that for ℓ_2 perturbations and isotropic features, since $w^\star = (\alpha^{\star 2} + \mu^{\star 2})^{1/2}$, the errors depend on the ratio $\alpha^\star/\mu^\star$. In fact, it can be seen that smaller values of the ratio lead to smaller adversarial errors. This leads to an interesting conclusion: *To find the hyper-parameter* $\varepsilon_{\rm tr}$ *that minimizes the adversarial error, it suffices to tune* $\varepsilon_{\rm tr}$ *to minimize the ratio* $\alpha^\star/\mu^\star$. A similar conclusion can be made for the case of ℓ_∞ perturbations, by noting from (10) that the adversarial error is characterized in a closed form in terms of $(\alpha^\star, \mu^\star, w^\star)$. In view of these observations, our sharp guarantees for the performance of adversarial training open the way to answering questions on the statistical limits and optimality of adversarial training, e.g., *how to optimally tune* $\varepsilon_{\rm tr}$*? How to optimally choose the loss function and what are the best minimum values of adversarial error achieved by the family of robust estimators? How do these answers depend on the adversary budget and/or the sampling ratio* δ*?* Fundamental questions of this nature have been recently addressed in the non-adversarial case based on the corresponding saddle point equations for standard ERM, e.g., [2], [9], [26], [27], [43], [44]. Theorems 1 and 2 are the first steps toward such extensions to the adversarial settings.

VI. Conclusion

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 GLM and GMM. Numerical simulations validate theoretical predictions even for relatively small problem dimensions and demonstrate the role of all problem parameters 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 [34] and neural tangent kernel [22] models. One other natural question is considering attacks other than ℓ_q-norm attacks considered in this article.

REFERENCES

[1] Z. Allen-Zhu and Y. Li, "Feature purification: How adversarial training performs robust deep learning," in *Proc. IEEE 62nd Annu. Symp. Found. Comput. Sci. (FOCS)*, Feb. 2022, pp. 977–988.

[2] D. Bean, P. J. Bickel, N. El Karoui, and B. Yu, "Optimal M-estimation in high-dimensional regression," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 36, pp. 14563–14568, Sep. 2013.

[3] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 32, pp. 15849–15854, Aug. 2019.

[4] M. Belkin, D. Hsu, and J. Xu, "Two models of double descent for weak features," *SIAM J. Math. Data Sci.*, vol. 2, no. 4, pp. 1167–1180, Jan. 2020.

[5] A. N. Bhagoji, D. Cullina, and P. Mittal, "Lower bounds on adversarial robustness from optimal transport," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 7498–7510.

[6] E. J. Candès and P. Sur, "The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression," *Ann. Statist.*, vol. 48, no. 1, pp. 27–42, Feb. 2020.

[7] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. S. Liang, "Unlabeled data improves adversarial robustness," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[8] M. Celentano and A. Montanari, "CAD: Debiasing the LASSO with inaccurate covariate model," 2021, *arXiv:2107.14172*.

[9] M. Celentano and A. Montanari, "Fundamental barriers to high-dimensional regression with convex penalties," *Ann. Statist.*, vol. 50, no. 1, pp. 170–196, Feb. 2022.

[10] M. Celentano, A. Montanari, and Y. Wei, "The LASSO with general Gaussian designs with applications to hypothesis testing," 2020, *arXiv:2007.13716*.

[11] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," 2018, *arXiv:1810.00069*.

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

[13] C. Dan, Y. Wei, and P. Ravikumar, "Sharp statistical guaratees for adversarially robust Gaussian classification," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 2345–2355.

[14] Z. Deng, A. Kammoun, and C. Thrampoulidis, "A model of double descent for high-dimensional logistic regression," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 4267–4271.

[15] O. Dhifallah and Y. Lu, "On the inherent regularization effects of noise injection during training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2665–2675.

[16] O. Dhifallah and Y. M. Lu, "A precise performance analysis of learning with random features," 2020, *arXiv:2008.11904*.

[17] E. Dobriban, H. Hassani, D. Hong, and A. Robey, "Provable tradeoffs in adversarially robust classification," 2020, *arXiv:2006.05161*.

[18] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, "Limitations of lazy training of two-layers neural network," in *Proc. NIPS*, 2019.

[19] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, "Modeling the influence of data structure on learning in neural networks: The hidden manifold model," *Phys. Rev. X*, vol. 10, no. 4, 2020, Art. no. 041044.

[20] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[21] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," *Ann. Statist.*, vol. 50, no. 2, pp. 949–986, Apr. 2022.

[22] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[23] A. Javanmard and M. Soltanolkotabi, "Precise statistical analysis of classification accuracies for adversarial training," *Ann. Statist.*, vol. 50, no. 4, pp. 2127–2156, Aug. 2022.

[24] A. Javanmard, M. Soltanolkotabi, and H. Hassani, "Precise tradeoffs in adversarial training for linear regression," in *Proc. Conf. Learn. Theory*, 2020, pp. 2034–2078.

[25] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[26] X. Mai and Z. Liao, "High dimensional classification via regularized and unregularized empirical risk minimization: Precise error and optimal loss," 2019, *arXiv:1905.13742*.

[27] X. Mai, Z. Liao, and R. Couillet, "A large scale analysis of logistic regression: Asymptotic performance and new insights," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3357–3361.

[28] S. Mei and A. Montanari, "The generalization error of random features regression: Precise asymptotics and the double descent curve," *Commun. Pure Appl. Math.*, vol. 75, no. 4, pp. 667–766, Apr. 2022.

[29] Y. Min, L. Chen, and A. Karbasi, "The curious case of adversarially robust models: More data can help, double descend, or hurt generalization," in *Proc. Uncertainty Artif. Intell.*, 2021, pp. 129–139.

[30] A. Montanari, F. Ruan, Y. Sohn, and J. Yan, "The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime," 2019, *arXiv:1911.01544*.

[31] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2574–2582.

[32] A. Raghunathan, S. M. Xie, F. Yang, J. Duchi, and P. Liang, "Understanding and mitigating the tradeoff between robustness and accuracy," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7909–7919.

[33] A. Raghunathan, S. M. Xie, F. Yang, J. C. Duchi, and P. Liang, "Adversarial training can hurt generalization," in *Proc. ICML Workshop Identifying Understand. Deep Learn. Phenomena*, 2019.

[34] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 20, 2007.

[35] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, vol. 317. Springer, 2009.

[36] F. Salehi, E. Abbasi, and B. Hassibi, "The impact of regularization on high-dimensional logistic regression," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[37] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry, "Adversarially robust generalization requires more data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5014–5026.

[38] S. H. Silva and P. Najafirad, "Opportunities and challenges in deep learning adversarial robustness: A survey," 2020, *arXiv:2007.00753*.

[39] M. Stojnic, "Various thresholds for ℓ_1-optimization in compressed sensing," 2009, *arXiv:0907.3666*.

[40] M. Stojnic, "A framework to characterize performance of LASSO algorithms," 2013, *arXiv:1303.7291*.

[41] P. Sur and E. J. Candès, "A modern maximum-likelihood theory for high-dimensional logistic regression," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 29, 2019, Art. no. 201810420.

[42] C. Szegedy et al., "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.

[43] H. Taheri, R. Pedarsani, and C. Thrampoulidis, "Sharp asymptotics and optimal performance for inference in binary models," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 3739–3749.

[44] H. Taheri, R. Pedarsani, and C. Thrampoulidis, "Fundamental limits of ridge-regularized empirical risk minimization in high dimensions," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 2773–2781.

[45] H. Taheri, R. Pedarsani, and C. Thrampoulidis, "Asymptotic behavior of adversarial training in binary linear classification," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2022, pp. 127–132.

[46] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "Precise error analysis of regularized M-estimators in high dimensions," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5592–5628, Aug. 2018.

[47] C. Thrampoulidis, S. Oymak, and B. Hassibi, "Regularized linear regression: A precise analysis of the estimation error," in *Proc. 28th Conf. Learn. Theory*, 2015, pp. 1683–1709.

[48] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *Proc. Int. Conf. Learn. Represent.*, 2019.

[49] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. Int. Conf. Mach. Learn.*, 2019.

Hossein Taheri received the B.Sc. degree in electrical engineering and mathematics from the Sharif University of Technology, Tehran, Iran, in 2018. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of California at Santa Barbara, Santa Barbara, CA, USA.

His main area of research is on statistical learning.

Ramtin Pedarsani (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Tehran, Tehran, Iran, in 2009, the M.Sc. degree in communication systems from the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2011, and the Ph.D. degree from the University of California at Berkeley, Berkeley, CA, USA, in 2015.

He is an Associate Professor with the ECE Department, University of California at Santa Barbara, Santa Barbara, CA, USA. His research interests include machine learning, information and coding theory, networks, and transportation systems.

Dr. Pedarsani was a recipient of the Communications Society and Information Theory Society Joint Paper Award in 2020, the Best Paper Award at the IEEE International Conference on Communications in 2014, and the NSF CRII Award in 2017.

Christos Thrampoulidis received the Diploma degree in ECE from the University of Patras, Patras, Greece, in 2011, and the M.Sc. and Ph.D. degrees in electrical engineering with a minor in applied and computational mathematics from the California Institute of Technology, Pasadena, CA, USA, in 2012 and 2016, respectively.

He was an Assistant Professor at the University of California at Santa Barbara, Santa Barbara, CA, USA, and a Post-Doctoral Researcher at Massachusetts Institute of Technology, Cambridge, MA, USA. He is an Assistant Professor with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada. His research is on high-dimensional estimation, machine learning, and optimization.