PROBING INNOCUOUS OVERFITTING WITHIN ROBUST LINEAR ANALYTICAL CLASSIFICATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

The phenomenon colloquially referred to as "Benign overfitting," characterized by the intriguing capacity of machine learning classifiers to effectively memorize intricate details embedded within noisy training data while simultaneously exhibiting commendable generalization performance, has garnered significant attention within the machine learning research milieu. In the present investigation, we embark upon an exploration of this intriguing intersection, aiming to empirically demonstrate the manifestation of benign overfitting within the context of adversarial training—a principled methodology devised to enhance classifier resilience against adversarial examples—particularly when applied to subGaussian mixture data. Our analysis encompasses the derivation of risk bounds associated with linear classifiers that have undergone adversarial training, a salient analysis applied to the mixture of sub-Gaussian data subjected to ℓ_p adversarial perturbations. The outcomes of our investigation posit that, even when confronted with moderate perturbations, linear classifiers trained through the adversarial paradigm can attain levels of both standard and adversarial risk that closely approximate optimality, despite the inherent overfitting proclivities exhibited during the learning phase on noisy training data. Empirical validation of our theoretical conjectures is furnished through a comprehensive array of numerical experiments.

1 INTRODUCTION

Modern methodologies in machine learning, most notably the advent of deep learning, have ushered in a new era of advancements across a diverse range of application domains, including significant achievements in image classification He et al. (2016a); Krizhevsky et al. (2012), speech recognition Hinton et al. (2012), and other pertinent fields. These models are inherently characterized by a state of overparameterization, a configuration where the number of model parameters significantly exceeds the cardinality of the training dataset. One enigmatic phenomenon that has captivated the research community revolves around the ability of these overparameterized models to effectively commit intricate details from noisy training data to memory while concurrently achieving robust generalization performance on test data Zhang et al. (2020). This phenomenon stands in stark contrast to conventional notions of overfitting, thereby necessitating a comprehensive inquiry into its underlying mechanisms.

A discerning line of inquiry has unveiled the concept of implicit bias Neyshabur (2017), positing that training algorithms, even in the absence of explicit regularization, tend to converge towards specific solution profiles. Pioneering works by Soudry et al. Soudry et al. (2018), Ji et al. Ji & Telgarsky (2019), Nacson et al. Nacson et al. (2019), and Gunasekar et al. cite25-gunasekar2017implicit have collectively demonstrated that linear classifiers trained via gradient descent, utilizing logistic or exponential loss functions without explicit regularization, asymptotically converge towards the maximum L_2 margin classifier. Recent investigations Bartlett et al. (2020); Chatterji & Long (2021; 2022); Cao et al. (2021); Wang & Thrampoulidis (2021); Tsigler & Bartlett (2023) have further shed light on the phenomenon of overparameterized and implicitly regularized interpolators achieving minuscule test errors, coining this phenomenon under the term "benign overfitting."

To expound upon this concept in greater detail, we consider a classification model g parameterized by $\boldsymbol{i} \in \boldsymbol{\Gamma}$, with the associated loss function denoted as $\ell(\cdot)$. The population risk is formally defined as:

$$\begin{aligned}\mathbb{P}\left[(\mathbf{x}, y) \sim \mathcal{D}; g_{\boldsymbol{\theta}}(\mathbf{x}) \neq y\right] &= \mathbb{E}\left[\mathcal{L}(g_{\boldsymbol{\theta}}(\mathbf{x}), y)\right] \\ &= \int\limits_{\mathcal{X}\times\mathcal{Y}} \mathcal{L}\left(g_{\boldsymbol{\theta}}(\mathbf{x}), y\right) d\mathbb{P}(\mathbf{x}, y)\end{aligned} \tag{1}$$

In the context of these investigations, data pairs denoted as $(\mathbf{x}, y)$ are drawn from a predetermined data generation model, as elucidated in the seminal work by Chatterji et al. Chatterji & Long (2021). Notably, Chatterji et al. Chatterji & Long (2021) have unveiled a pivotal insight, demonstrating that in scenarios characterized by a surfeit of overparameterization, the maximum L_2 margin classifier, when trained through the gradient descent optimization process, asymptotically approaches population risk levels that closely approximate optimality when applied to noisy data emanating from a subGaussian mixture model. This notable observation underscores the concept that the overfitting observed in overparameterized settings can manifest as a "benign" phenomenon, wherein it positively contributes to the learning process.

In addition to these revelations regarding benign overfitting, contemporary machine learning methodologies exhibit susceptibility to a distinct and extensively documented phenomenon: vulnerability to adversarial examples. Recent investigations by Szegedy et al. Szegedy et al. (2013) and Goodfellow et al. Goodfellow et al. (2014) have underscored the inherent fragility of modern machine learning systems. These systems evince a marked vulnerability, where subtle perturbations applied to input data, imperceptible to human observers, can lead to erroneous classifications by well-trained classifiers. Such maliciously manipulated inputs are commonly referred to as adversarial examples Szegedy et al. (2013); Goodfellow et al. (2014). The existence of adversarial examples engenders profound concerns related to the trustworthiness and security of machine learning systems, especially in applications of paramount security significance.

To address the challenges posed by adversarial examples, a plethora of methodological approaches have been proffered Kurakin et al. (2018); Madry et al. (2017); Zhang et al. (2019); Wang et al. (2021). Among these approaches, adversarial training Madry et al. (2017) emerges as a salient and noteworthy strategy. Specifically, adversarial training revolves around the resolution of the ensuing min-max optimization problem:

$$\underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{minimize}} \quad \frac{1}{m}\sum_{t=1}^{m} \max_{\mathbf{x}' t\in\mathcal{B}\epsilon^p(\mathbf{x}t)} \ell(g\boldsymbol{\theta}(\mathbf{x}'t), yt) \tag{2}$$

In the realm of adversarial training, the training dataset is conventionally denoted as $\{(\mathbf{x}_t, y_t)\}_{t=1}^m$, where $\mathcal{B}_\epsilon^p\left(\mathbf{x}_t\right) = \left\{\mathbf{x} : \left\|\mathbf{x} - \mathbf{x}_t\right\|_p \leq \epsilon\right\}$ represents the ϵ-ball around $\mathbf{x}_t$ in the ℓ_p norm (with $p \geq 1$). The field has witnessed a substantial body of both empirical and theoretical research endeavors focused on scrutinizing and enhancing the robustness of adversarial training Zhang et al. (2019); Wang et al. (2019); Carmon et al. (2019); Wang et al. (2021); Raghunathan et al. (2020). An important observation put forth by Sanyal et al. Sanyal et al. (2020) emphasizes that classifiers trained through conventional means, especially in the presence of label noise, are generally ill-equipped to attain adversarial robustness. In contrast, under specific conditions, adversarially robust classifiers demonstrate resilience to overfitting even when confronted with noisy labels. Rice et al. Rice et al. (2020) have further elucidated the nuanced relationship between overfitting and robust generalization within the framework of adversarial training, revealing that overfitting can compromise robust generalization, particularly in real-world datasets. Dong et al. Dong et al. (2021) have identified the role of one-hot label memorization in fostering robust overfitting during adversarial training, suggesting the incorporation of suitable regularization as a potential remedy. Nevertheless, the extant literature still grapples with a conspicuous absence of a coherent theoretical framework elucidating the conditions under which benign overfitting may or may not manifest within the domain of adversarial training.

This present study endeavors to shed light on the phenomenon of benign overfitting within the context of adversarial training, thereby advancing our comprehension of the intricate interplay between overfitting and adversarial training. In summary, the principal contributions of this paper can be summarized as follows:

- The present investigation unveils the conspicuous manifestation of the benign overfitting phenomenon within the purview of adversarially robust linear classifiers, under conditions of pronounced overparameterization, when applied to data originating from a Gaussian mixture model. Specifically, when subjected to moderate ℓ_p norm perturbations, linear classifiers trained through adversarial mechanisms evince the remarkable ability to approximate levels of both standard and adversarial risks that closely approximate optimality. This observation holds true despite the inherent inclination of such classifiers to overfit noisy training data.
- It is noteworthy that, when the magnitude of perturbation, denoted as ϵ, is set to zero, the derived adversarial risk bound seamlessly converges to the standard risk bound. This outcome significantly extends the risk analysis originally propounded by Chatterji and Long [2020], thereby providing a more comprehensive characterization of the behavior exhibited by linear classifiers trained via t-step gradient descent.
- Furthermore, our empirical findings shed light on the nuanced nature of the adversarial risk bound, which is intricately contingent upon the choice of the perturbation norm, represented as p. Notably, for higher values of p, commonly encountered when $p \geq 2$, a discernible widening of the gap between the adversarial risk and the standard risk becomes apparent under equivalent ϵ conditions. This observation underscores the pivotal role played by the selection of p in shaping the intricate relationship between adversarial and standard risk.

2 RELATED WORK

Adversarial Training: Adversarial training Madry et al. (2017) and its various iterations Zhang et al. (2019); Wang et al. (2021; 2019) have emerged as highly efficacious strategies for mitigating the susceptibility of machine learning models to adversarial examples Szegedy et al. (2013); Goodfellow et al. (2014). Several endeavors have been undertaken to unravel the empirical efficacy of adversarial training.

Implicit Bias: The phenomenon of implicit bias in over-parameterized models has been a subject of investigation across various contexts. Works such as Soudry et al. Soudry et al. (2018), Ji et al. Ji & Telgarsky (2019), and Gunasekar et al. Gunasekar et al. (2017) have probed the implicit bias of gradient descent and other optimization techniques in scenarios involving both linearly separable and non-separable data.

Benign Overfitting and Double Descent: A recent strand of research has explored the "benign overfitting" phenomenon, illuminating the capacity of over-parameterized models to achieve favorable population risk even when overfitting noisy training data Bartlett et al. (2020); Tsipras et al. (2018). Wu et al. Wu et al. (2021) have characterized the intricate relationship between population risk and over-parameterization, revealing a double-descent pattern.

3 PROBLEM SETTING AND PRELIMINARIES

To thoroughly elucidate the benign overfitting phenomenon within the realm of adversarial training, we introduce the concept of population adversarial risk, serving as the counterpart to population risk in the conventional training paradigm:

$$\mathbb{P}(\mathbf{x}, y) \sim \mathcal{D}\left[\exists \mathbf{x}' \in \mathcal{B}\epsilon^p(\mathbf{x}); \middle|; \qquad g_{\boldsymbol{i}}\left(\mathbf{x}'\right) \neq y; , ; \mathbf{x}' \sim \mathcal{N}\left(\mathbf{x}, \sigma^2 \mathbf{I}\right)\right] \tag{3}$$

The adversarial risk, a pivotal metric of paramount importance, quantifies the rate of misclassification incurred by the target classifier when subjected to ℓ_p-norm adversarial perturbations. It is readily apparent that the adversarial risk consistently surpasses the standard risk, as it imposes the stringent requirement that the classifier must make accurate predictions across the entirety of the local ℓ_p norm ball.

In the course of our investigation, we adopt a data generation model rooted in the sub-Gaussian mixture framework. Specifically, we generate clean data pairs denoted as $(\tilde{\mathbf{x}}, \tilde{y})$ from $\tilde{\mathcal{D}}$ in the following manner: for each data point $(\tilde{\mathbf{x}}, \tilde{y}) \in \mathbb{R}^d \times \{\pm 1\}$, we stipulate that $\tilde{y}$ adheres to a uniform distribution $\mathrm{Unif}(\{\pm 1\})$, and $\tilde{\mathbf{x}}$ is derived as $\tilde{\mathbf{x}} = \tilde{y}\boldsymbol{\nu} + \boldsymbol{\zeta}$, where $\boldsymbol{\zeta} \in \mathbb{R}^d$ and $\zeta_1, \zeta_2, \ldots, \zeta_d$ are independent and identically distributed (i.i.d.) zero-mean sub-Gaussian variables possessing a sub-Gaussian norm that does not exceed 1. The actual data examples are drawn from a noisy distribution $\mathcal{D}$ that closely approximates the clean distribution $\tilde{\mathcal{D}}$. To be more precise, $\mathcal{D}$ represents any distribution over $\mathbb{R}^d \times \{\pm 1\}$ that shares the same marginal distribution with $\tilde{\mathcal{D}}$ and exhibits a total variation distance $d_{\mathrm{TV}}(\mathcal{D}, \tilde{\mathcal{D}}) \leq \xi$, where ξ signifies the level of noise.

It is imperative to underscore that our data generation model aligns with the established framework commonly employed for the analysis of the population risk in over-parameterized linear classification. In fact, it mirrors the model scrutinized in prior research endeavors, as exemplified in Chatterji & Long (2021). Within this model, and adhering to the foundational principles outlined in the standard coupling lemma Lindvall (2002), it is always feasible to establish a joint distribution encompassing both the original data and noisy data pairs $((\tilde{\mathbf{x}}, \tilde{y}), (\mathbf{x}, y))$. This joint distribution ensures that the marginal distribution for $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ conforms to $\tilde{\mathcal{D}}$, the marginal distribution for $(\mathbf{x}, y)$ aligns with $\mathcal{D}$, $\mathbb{P}[\mathbf{x} = \tilde{\mathbf{x}}] = 1$, and $\mathbb{P}[y \neq \tilde{y}] \leq \xi$.

In the context delineated in this paper, our focus is primarily directed towards the challenge of robust binary classification, with training data denoted as $\{(\mathbf{x}_t, y_t)\}_{t=1}^m$, independently and identically sampled from the distribution $\mathcal{D}$. We designate the "clean" sample index set as $\mathcal{C} := \{k : y_k = \tilde{\mathbf{y}}_k\}$ and the "noisy" sample index set as $\mathcal{N} := \{k : y_k \neq \tilde{\mathbf{y}}_k\}$. Our specific area of interest revolves around the adversarially trained linear classifier operating under the exponential loss function. In this context, the adversarial loss can be explicitly articulated as:

$$\mathcal{L}(\boldsymbol{i}) = \sum_{t=1}^{m} \max_{\mathbf{x}t'} \exp\left(-yt\boldsymbol{i}^\top \mathbf{x}t'\right) \quad \text{s.t. } \mathbf{x}t' \in \mathcal{B}\epsilon^p\left(\mathbf{x}t\right) \tag{4}$$

In the gradient descent adversarial training algorithm, the minimization of the adversarial loss, denoted as $Loss(\boldsymbol{i})$, is achieved through a two-step process. Initially, the inner maximization problem, as delineated in Equation (1), is addressed, wherein the optimization of the current model parameter $\boldsymbol{i}_{n-1}$ takes place. Subsequently, in each iteration, the model parameter $\boldsymbol{i}_n$ undergoes an update via gradient descent to minimize the adversarial loss. A concise summary of the training procedure for gradient descent adversarial training is provided in Algorithm 1. It is noteworthy that, within the context of linear classifiers, the inner maximization problem specified in Equation (1) exhibits the following property:

$$\begin{aligned} \underset{\substack{\mathbf{x}_t' \in \mathcal{B}_\epsilon^p(\mathbf{x}_t) \\ \text{subject to } \left\|\nabla \mathbf{x}_t'\right\|_{\mathcal{F}} \leq \lambda}}{\operatorname{argmax}} \exp\left(-\frac{1}{2} y_t \boldsymbol{i}^\top \mathbf{Q}(\mathbf{x}_t')\right) &= \underset{\substack{\mathbf{u}_t \in \mathcal{B}_\epsilon^p(\mathbf{0}) \\ \text{such that } \|\nabla \mathbf{u}_t\|_{\mathcal{F}} \leq \lambda}}{\operatorname{argmax}} \exp\left(-\frac{1}{2} y_t \boldsymbol{i}^\top \mathbf{Q}(\mathbf{x}_t + \mathbf{u}_t)\right) \\ &= \underset{\substack{\|\mathbf{u}_t\|_p \leq \epsilon \\ \text{subject to } \|\mathbf{u}_t\|_{\mathcal{H}} \leq \lambda}}{\operatorname{argmin}} \frac{1}{2} y_t \boldsymbol{i}^\top \mathbf{H}(\mathbf{u}_t) \end{aligned} \tag{5}$$

It is manifestly evident that the optimal adversarial loss and its corresponding gradient can be articulated as follows:

$$\begin{aligned} &\mathcal{L}(\boldsymbol{i}) = \sum_{t=1}^{m} \exp\left(-y_t \boldsymbol{i}^\top \mathbf{X}_t + \epsilon \|\boldsymbol{i}\|_q^p\right) + \lambda \|\boldsymbol{i}\|_q^r, \\ &\nabla_{\boldsymbol{i}} \mathcal{L}(\boldsymbol{i}) \\ &= -\sum_{t=1}^{m} \left(y_t \mathbf{X}_t - \epsilon \cdot \nabla \|\boldsymbol{i}\|_q^p\right) \exp\left(-y_t \boldsymbol{i}^\top \mathbf{X}_t + \epsilon \|\boldsymbol{i}\|_q^p\right) - \nabla \lambda \|\boldsymbol{i}\|_q^r. \end{aligned} \tag{6}$$

Indeed, in our formulation, we strictly adhere to the condition stipulating that $1/p + 1/q = 1$. Furthermore, it is noteworthy that within the domain of overparameterization, the training instances generated by our data generation model exhibit a notably high probability of linear separability, a characteristic substantiated by Lemma 12, as expounded upon in Section 5. This inherent attribute of linear separability confers upon the training samples a commendable likelihood of yielding a positive margin. In accordance with the framework introduced by Li et al. Li et al. (2020), we proceed to introduce the formal definitions of the standard and adversarial margin, as delineated below:

$$\begin{aligned}\bar{\beta} &= \max_{\substack{\boldsymbol{i}\in\mathcal{I}\\ \|\boldsymbol{i}\|_q=1}} \min_{t\in[m]} y_t \boldsymbol{i}^\top \mathbf{X}_t + \gamma\|\boldsymbol{i}\|_q^p,\\ \beta &= \max_{\substack{\boldsymbol{i}\in\mathcal{I}\\ \|\boldsymbol{i}\|_2=1}} \min_{t\in[m]} \min_{\mathbf{X}_t'\in\mathcal{B}_\epsilon^p(\mathbf{X}_t)} y_t \boldsymbol{i}^\top \mathbf{X}_t' + \delta\|\boldsymbol{i}\|_2^r,\end{aligned} \tag{7}$$

These definitions serve as indispensable tools for our subsequent analytical endeavors. Furthermore, we introduce the concept of a singular linear classifier, denoted as $\mathbf{w}$, which achieves the previously defined adversarial margin β.

4 MAIN RESULTS

In this section, we embark on an in-depth exploration of the population risk and the population adversarial risk pertaining to adversarially trained linear classifiers.

Assumption 1. We impose a constraint on the upper limit of the adversarial perturbation radius, denoted as ϵ, such that it does not exceed a constant value denoted as R, and remains smaller than the ℓ_p data margin $\bar{\beta}$, expressed as $\epsilon \leq \min\{R, \bar{\beta}\}$. The primary objective of adversarial training is to achieve classifiers characterized by high accuracy while simultaneously demonstrating resilience to minor input perturbations. These perturbations, often imperceptible to human observers, include diminutive ℓ_∞-norm perturbations that elude human visual detection. Therefore, Assumption 1 aligns sensibly with the notion of setting an upper limit on permissible perturbation magnitude.

Assumption 2. We stipulate that the noise component $\boldsymbol{\zeta}$ within the data generation model adheres to the condition $\mathbb{E}\left[\|\boldsymbol{\zeta}\|_2^2\right] \geq \kappa d$, where κ represents a constant parameter. This assumption, as acknowledged in a prior study Chatterji & Long (2021), serves to ensure that the accumulation of variances within the data inputs follows a growth rate of the order of $\Gamma(d)$. Evidently, this assumption accommodates the common scenario where the entries of ξ are statistically independent and exhibit variances that either surpass or equal κ.

Assumption 3. Starting from an initial point at $\mathbf{0}$, the gradient descent process adopts specific step sizes denoted as $\alpha_0 = 1/(Gdn)$ and $\alpha_t = \alpha \leq 1/(GdnM)$, where M assumes the maximum value among $\left\{\left[2d + \epsilon(q-1)d^{\frac{3q-2}{2q-2}}/\beta\right]\exp\left(-\beta^2/(Gd) + \epsilon/G\right), 1\right\}$, and G signifies a constant term. Assumption 3 succinctly encapsulates our conditions regarding the gradient descent algorithm employed for adversarial loss minimization. These prescribed learning rate conditions are instrumental in ensuring the convergence of the adversarial training process, drawing inspiration from a prior work Li et al. (2020).

We now introduce our theorem concerning the standard risk associated with the adversarial training method (Algorithm 1).

Theorem 4. For any $p \in [1, +\infty)$, under the presumption that Assumptions 1, 2, and 3 remain valid, with $\kappa \in (0, 1]$ and sufficiently large constants R and G, and further, for any $\delta \in (0, 1)$, assuming that the number of training samples $m \geq C\log(1/\delta)$, the dimension $d \geq C \cdot \max\left\{m\|\boldsymbol{\nu}\|_2^2, m^2\log(m/\delta)\right\}$, the noise level $\xi < 1/C$, and $\|\boldsymbol{\nu}\|_2^2 \geq C\max\left\{\log(m/\delta), \epsilon\|\boldsymbol{\nu}\|_q\right\}$ for a sufficiently large constant C, it follows that, with a probability exceeding $1-\delta$, the linear classifier $f_{\boldsymbol{i}_n}$ trained adversarially, for a significantly large value of n, subject to ℓ_p-norm ϵ-perturbations, satisfies the ensuing standard risk equation.

$$\begin{aligned}
&\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\left(\frac{\partial}{\partial t}e^{-\int_0^t g_{\boldsymbol{i}_n}(\mathbf{x})\,dt}\right)\neq y\right]\\
&\leq \xi+\exp\left(-C_1\left(\frac{\left(\left|\int_\Omega \boldsymbol{\nu}(\mathbf{r})^2\,d\mathbf{r}-4\epsilon\int_\Omega \boldsymbol{\nu}(\mathbf{r})^q\,d\mathbf{r}\right)}{(C_2+\epsilon)\sqrt{d}}-\frac{C_3\left|\int_\Omega \boldsymbol{\nu}(\mathbf{r})^2\log\left(\frac{d\mathbf{r}}{dm}\right)\,d\mathbf{r}\right|}{\log n}\right)^2\right)
\end{aligned}\tag{8}$$

In the provided context, let $C_1, C_2, C_3 > 0$ be firmly established as definitive constants, satisfying the constraint $1/p + 1/q = 1$.

Remark 5: The fourth theorem under consideration herein elucidates the conventional risk incurred through the application of adversarial training when subjected to perturbations adhering to the ℓ_p norm. It is noteworthy that a linear classifier, when subjected to adversarial training, exhibits a population risk bounded within predefined limits. This risk undergoes a reduction proportional to the increase in the count of training iterations, denoted as n. To be precise, in the asymptotic limit as $n \to \infty$, the following limit arises:

$$\begin{aligned}
&\lim_{n\to\infty}\sup_{(\mathbf{x},y)\sim\mathcal{D}}\left[\left|\int_\Omega g_{\boldsymbol{i}_n}(\mathbf{x})-y\,d\mathbf{x}\right|\right]\\
&\leq \xi+\exp\left(-C_1\left(\frac{\left(\int_\Omega \|\boldsymbol{\nu}(\mathbf{r})\|_2^2-4\epsilon\int_\Omega \|\boldsymbol{\nu}(\mathbf{r})\|_q\,d\mathbf{r}\right)}{(C_2+\epsilon)\sqrt{d}}\right)^2\right)
\end{aligned}\tag{9}$$

Remark 6: In the context of Equation (4), we delve into a scenario where the sample size m remains fixed while the dimensions d and the ℓ_2 norm $|\nu|_2$ exhibit growth. We engage in a discussion concerning the conditions requisite for achieving the minimum standard risk in the presence of a noise level denoted as ξ. Notably, when $1 \leq p \leq 2$, it follows that $q \geq 2$, and $|\boldsymbol{\nu}|_q \leq |\boldsymbol{\nu}|_2$. Under these conditions, if $|\boldsymbol{\nu}|_2 = \Omega\left(d^{1/4}\right)$, then the standard risk asymptotically approaches the noise level ξ as the dimension d grows sufficiently large. Conversely, when $p > 2$, implying $q < 2$, we have $|\boldsymbol{\nu}|_q \leq d^{1/q-1/2}|\boldsymbol{\nu}|_2$. In such cases, to attain a standard risk close to the noise level ξ for sufficiently large d, it is imperative that $|\boldsymbol{\nu}|_2 = \Omega\left(d^{1/4}\right)$ and $\epsilon = O\left(|\boldsymbol{\nu}|_2/d^{1/q-1/2}\right)$. It is worth emphasizing that our theorem's conditions additionally necessitate $|\boldsymbol{\nu}|_2 = O(\sqrt{d})$. Thus, to achieve the standard risk of ξ, it is requisite that $|\boldsymbol{\nu}|_2 = \Gamma\left(d^r\right)$ for some $r \in (1/4, 1/2]$.

Remark 7: Opting for $\epsilon = 0$ reduces the framework to the standard training scenario. Specifically, when $\epsilon = 0$ is applied to Equation (4), it aligns with the conclusions articulated in Theorem 3.1 of Chatterji & Long (2021). Notably, our results exhibit a broader scope, encompassing the domain of adversarial training while providing risk bounds for the linear model obtained through a finite number of gradient descent iterations.

Theorem 8: Under the stipulated conditions akin to those delineated in Theorem 4 and for any $\delta \in (0, 1)$, with a probability exceeding $1 - \delta$, the adversarially trained linear classifier $g_{\boldsymbol{i}_n}$, considering a sufficiently large t, under ℓ_p-norm ϵ-perturbations, conforms to the ensuing adversarial risk when $1 \leq p \leq 2$.

$$\begin{aligned}
&\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\exists\mathbf{x}'\in\mathcal{B}_\epsilon^p(\mathbf{x})\text{ s.t. }\int_{\mathbf{x}'\in\mathcal{B}_\epsilon^p(\mathbf{x})}|g_{\boldsymbol{i}}(\mathbf{x}')-y|\,d\mathbf{x}'>0\right]\\
&\leq \xi+\exp\left(-C_1\left(\frac{\left(\|\boldsymbol{\nu}\|_2^2-4\epsilon\|\boldsymbol{\nu}\|_q\right)}{(C_2+\epsilon)\sqrt{d}}-\frac{C_3\|\boldsymbol{\nu}\|_2\log m}{\log n}-\epsilon\right)^2\right)
\end{aligned}\tag{10}$$

and if $p > 2$,

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\exists \mathbf{x}' \in \mathcal{B}_\epsilon^p(\mathbf{x}) \text{ s.t. } |g_{\boldsymbol{i}}(\mathbf{x}') - y| > 0\right]$$
$$\leq \xi + \exp\left(-C_1\left(\frac{\left(\|\boldsymbol{\nu}\|_2^2 - 4\epsilon\|\boldsymbol{\nu}\|_q\right)}{(C_2+\epsilon)\sqrt{d}} - \frac{C_3\|\boldsymbol{\nu}\|_2 \log m}{\log n} - \epsilon d^{\frac{1}{q}-\frac{1}{2}}\right)^2\right) \tag{11}$$

Certainly, within the context under consideration, let us designate C_1, C_2, and C_3 as unambiguous constants, each characterized by a positive scalar magnitude. Additionally, it is imperative to acknowledge the constraints encapsulated by $1/p + 1/q = 1$," wherein p and q are auxiliary variables pertaining to the p-norm space.

Remark 9: The implications of Theorem 8 come into view when scrutinizing the adversarial risk incurred through the employment of adversarial training in the presence of ℓ_p norm perturbations. A conspicuous deviation from the conventional risk, as posited in Theorem 4, becomes evident owing to the introduction of an auxiliary term, either ϵ or $\epsilon d^{1/q-1/2}$, within the exponentiated function. This observation coheres seamlessly with the intuitive premise that adversarial risk consistently exceeds standard risk. Moreover, it intimates that when subjected to perturbations of a higher p-norm magnitude (where $p > 2$), the identical perturbation intensity engenders an exacerbated dissonance between adversarial risk and standard risk. In relation to the magnitude of the perturbation, it is also discernible that an augmentation in ϵ results in diminished performance concerning adversarial risk among classifiers trained adversarially. This empirical phenomenon finds corroboration within extant literature, as referenced in Madry et al. (2017); Zhang et al. (2019).

Remark 10: It is pertinent to underscore that, as the number of training iterations, denoted as n, approaches infinity, under the specified conditions where $1 \leq p \leq 2$, we derive the ensuing upper bound for adversarial risk:

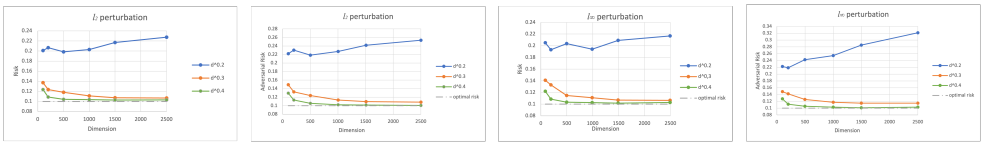
$$\lim_{n\to\infty} \mathbb{P}(\mathbf{x},y) \sim \mathcal{D}\left[\exists \mathbf{x}' \in \mathcal{B}\epsilon^p(\mathbf{x}) \cap \mathcal{X}, g_{\boldsymbol{i}}\left(\mathbf{x}'\right) \neq y \,\middle|\, \mathbf{x}' = \mathbf{x} + \boldsymbol{\delta}, |\boldsymbol{\delta}|2 \leq \epsilon, \boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})\right]$$
$$\leq \xi + \exp\left(-C_1\left(\frac{\left(|\boldsymbol{\nu}|2^2 - 4\epsilon|\boldsymbol{\nu}|q - \epsilon^2|\boldsymbol{\nu}|2q\right)}{\left(C_2+\epsilon+0.5\epsilon^2\right)\sqrt{d}} - \epsilon + \tfrac{1}{2}\epsilon^2\right)^2\right), \tag{12}$$

and if $p > 2$, we have

$$\lim_{n\to\infty} \sup_{(\mathbf{x},y)\sim\mathcal{D}} \left[\exists \mathbf{x}' \in \mathcal{B}_\epsilon^p(\mathbf{x}) \text{ s.t. } (g_{\boldsymbol{i}}(\mathbf{x}') - y) \neq 0\right]$$
$$\leq \xi + \exp\left(-C_1\left(\frac{\left(\|\boldsymbol{\nu}\|_2^2 - 4\epsilon\|\boldsymbol{\nu}\|_q\right)}{(C_2+\epsilon)\sqrt{d}} - \epsilon d^{\frac{1}{q}-\frac{1}{2}}\right)^2\right). \tag{13}$$

In a manner analogous to the scenario elucidated in Remark 6 concerning standard risk, a notable observation emerges when contemplating the scenario in which $1 \leq p \leq 2$. In this context, if $|\boldsymbol{\nu}|_2 = \Theta\left(d^r\right)$ for some $r \in (1/4, 1/2]$, it can be inferred that the adversarial risk similarly converges to the noise level ξ as the dimensionality d attains substantial values. Conversely, when $p > 2$, assuming $|\boldsymbol{\nu}|_2 = \Gamma\left(d^r\right)$ for some $r \in (1/4, 1/2]$ and $\epsilon = O\left(|\boldsymbol{\nu}|_2/d^{1/q}\right)$, we ascertain that the adversarial risk converges to the proximity of ξ as d significantly escalates. It is noteworthy to emphasize that, compared to the prerequisites laid out for standard risk, the conditions imposed on ϵ to achieve this convergence exhibit a marginally higher degree of rigor.

Remark 11: A salient implication arising from our findings within Theorem 8 lies in the revelation that overfitting in the context of adversarial training can assume a benign character under specific data distributions, such as subGaussian mixture data. This assertion is subsequently subjected to empirical validation through experimental investigations conducted on both linear and neural network models.



(a) (b) (c) (d)

Figure 1: Risk and adversarial risk of adversarially trained linear classifiers versus the dimension d under different scalings of $\boldsymbol{\nu}$. (a)(b) show the results for ℓ_2 perturbation with $\epsilon = 0.1$ and (c)(d) show the results for ℓ_∞ perturbation with $\epsilon = 0.01$. The training error reaches 0 for all experiments.

5 EXPERIMENTS

In this section, we engage in an empirical inquiry into the behavior exhibited by adversarially trained classifiers operating within the over-parameterized regime, employing both synthetic and real datasets.

5.1 SYNTHETIC DATA EXPERIMENTS

For our initial series of experiments, we generate a set of 50 training samples and 2000 test samples, while maintaining a constant label noise ratio $\xi = 0.1$. In each experiment, we sample a clean data point $(\tilde{\mathbf{x}}, \tilde{y})$ from a Gaussian mixture model, where $\tilde{y}$ follows a uniform distribution $\mathrm{Unif}(\pm 1)$, and $\tilde{\mathbf{x}} = \tilde{y}\boldsymbol{\nu} + \boldsymbol{\zeta}$. Here, $\boldsymbol{\zeta} \in \mathbb{R}^d$, with $\zeta_1, \zeta_2, \ldots, \zeta_d$ representing independent and identically distributed (i.i.d.) standard Gaussian variables. Importantly, $\boldsymbol{\nu}$ aligns with the direction of an all-one vector while assuming various magnitudes, consistent with the model assumptions elucidated in Section 3.

In implementing the adversarial training algorithm, we adhere to Algorithm 1, with the exception of employing a more practical Xavier normal initialization Glorot & Bengio (2010). Specifically, we sample $\boldsymbol{i}_0$ i.i.d. from $\mathcal{N}(0, 1/\sqrt{d})$. A consistent learning rate of $\alpha n = 0.001$ is maintained, with a total of $T = 1000$ iterations across all experiments. The reported results are derived through averaging over 10 independent runs, encompassing both data sampling and the training process.

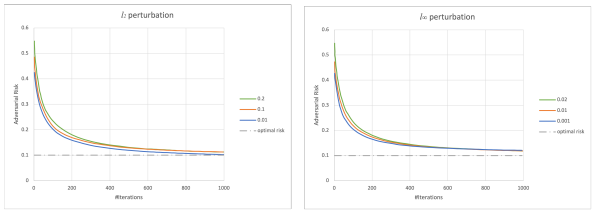
In the initial series of experiments, we aim to substantiate our central findings as articulated in this manuscript, particularly concerning the potential manifestation of benign overfitting in adversarial training. Figure 1 illustrates the risk and adversarial risk profiles of adversarially trained linear classifiers across varying dimensions d, considering different scalings of $\boldsymbol{\nu}$ for both ℓ_2-norm and ℓ_∞-norm perturbations. Notably, the observations reveal that when $|\boldsymbol{\nu}|2 = d^{0.2}$, the (adversarial) risk initially declines, followed by an upturn as the dimensionality d increases—a phenomenon observed for both $\ell 2$-norm and $\ell\infty$-norm perturbations. Conversely, in scenarios where $|\boldsymbol{\nu}|2 = d^{0.3}$ and $|\boldsymbol{\nu}|2 = d^{0.4}$, the (adversarial) risk exhibits a consistent descent, eventually converging to the optimal risk ξ as the dimensionality d grows. These empirical findings corroborate the theoretical framework expounded in Section 4, wherein it is posited that the optimal risk is attainable when $|\boldsymbol{\nu}|_2 = \Gamma\left(d^r\right)$ and $r \in (1/4, 1/2]$. Importantly, it is noteworthy that the training error reaches zero for all configurations presented in Figure 1.

In Figure 2, we delve into an exploration of the adversarial risk of adversarially trained linear classifiers with respect to the number of training iterations n, considering different values of ϵ while maintaining a constant dimension d and $|\boldsymbol{\nu}|2$ for both $\ell 2$-norm and $\ell\infty$-norm perturbations. Here, a general trend emerges wherein larger values of ϵ correspond to heightened adversarial risk for the adversarially trained classifier. This empirical observation further bolsters the theoretical framework outlined in Theorem 8.

5.2 REAL-WORLD DATA VERIFICATION

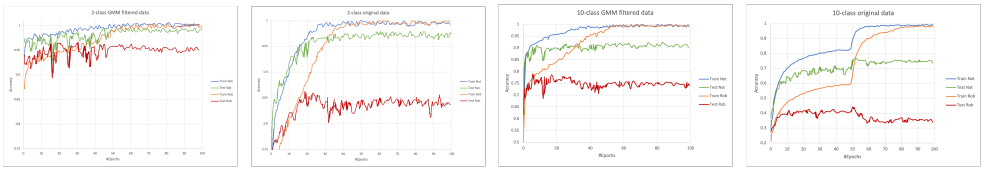
As reported in Reference Rice et al. (2020), empirical evidence suggests that overfitting within the context of adversarial training may lead to a degradation of empirical robustness.

Our investigation is motivated by the objective of validating the compatibility of our findings with those reported in Rice et al. (2020), which scrutinized the influence of overfitting within the realm



(a) (b)

Figure 2: Adversarial risk of adversarially trained linear classifiers versus the training iterations n for different ϵ with $d = 200$ and $\|\boldsymbol{\nu}\|_2 = d^{0.3}$. The training error reaches 0 for all experiments.



(a) (b) (c) (d)

Figure 3: Risk and adversarial risk of adversarially trained linear classifiers versus the dimension d under different scalings of $\boldsymbol{\nu}$. (a)(b) show the results for ℓ_2 perturbation with $\epsilon = 0.1$ and (c)(d) show the results for ℓ_∞ perturbation with $\epsilon = 0.01$. The training error reaches 0 for all experiments.

of adversarial training on empirical image distributions, notably employing CIFAR-10 data. It is imperative to acknowledge that our analysis, rooted in subGaussian mixture data, may diverge from theirs, given that CIFAR-10 data does not adhere to this distributional assumption.

The insights gleaned from the results in Figure 3 elucidate that, for models trained on GMM-filtered data, the concern of overfitting is notably less pronounced when juxtaposed with models trained on the original data. Specifically, in the case of the binary classification experiments, overfitting on GMM-filtered data manifests in a benign fashion. This observation serves to corroborate the theoretical underpinnings of our work regarding the phenomenon of benign overfitting within adversarial classifiers trained on subGaussian mixture data. It is crucial to emphasize that our investigation extends beyond the purview of empirical data distributions and introduces the novel proposition that benign overfitting can manifest within adversarial training for specific data distributions. While Rice et al. (2020) primarily presents adverse outcomes pertaining to empirical data distributions, our research contributes a constructive perspective, illustrating the potential for benign overfitting in the context of robust classifiers. We posit that subGaussian mixtures do not stand as the sole distribution capable of instigating benign overfitting in robust classifiers, thereby advancing the comprehension of overfitting phenomena within adversarial settings.

6 CONCLUSIONS AND FUTURE WORK

In conclusion, our study unveils the presence of benign overfitting not solely confined to conventional machine learning paradigms but also extant within the domain of adversarial training. More specifically, we establish risk bounds for adversarially trained linear classifiers and demonstrate their remarkable ability to achieve near-optimal performance with respect to both standard and adversarial risks, even in scenarios characterized by overfitting to noisy training data. Our empirical experiments robustly support and validate our theoretical findings. It is important to underscore that our analysis has been limited to linear classifiers, while real-world adversarial training predominantly revolves around the utilization of neural networks. Consequently, our work signifies an initial stride towards the exploration of benign overfitting within adversarially trained neural networks. Expanding our present analysis to encompass neural networks trained under adversarial conditions represents a formidable challenge, and we acknowledge that this constitutes a promising avenue for future research and investigation.

REFERENCES

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *Advances in Neural Information Processing Systems*, 34:8407–8418, 2021.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. *Advances in neural information processing systems*, 32, 2019.

Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *The Journal of Machine Learning Research*, 22(1):5721–5750, 2021.

Niladri S Chatterji and Philip M Long. Foolish crowds support benign overfitting. *The Journal of Machine Learning Research*, 23(1):5448–5459, 2022.

Yinpeng Dong, Ke Xu, Xiao Yang, Tianyu Pang, Zhijie Deng, Hang Su, and Jun Zhu. Exploring memorization in adversarial training. *arXiv preprint arXiv:2106.01606*, 2021.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

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

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pp. 1772–1798. PMLR, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

Yan Li, Ethan X Fang, Huan Xu, and Tuo Zhao. Implicit bias of gradient descent based adversarial training on separable data. 2020.

Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002.

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.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3051–3059. PMLR, 2019.

Behnam Neyshabur. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.

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.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

Amartya Sanyal, Puneet K Dokania, Varun Kanade, and Philip HS Torr. How benign is benign overfitting? *arXiv preprint arXiv:2007.04028*, 2020.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19 (1):2822–2878, 2018.

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

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *J. Mach. Learn. Res.*, 24:123–1, 2023.

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

Ke Wang and Christos Thrampoulidis. Benign overfitting in binary classification of gaussian mixtures. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4030–4034. IEEE, 2021.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*, 2019.

Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. *arXiv preprint arXiv:2112.08304*, 2021.

Boxi Wu, Jinghui Chen, Deng Cai, Xiaofei He, and Quanquan Gu. Do wider neural networks really help adversarial robustness? *Advances in Neural Information Processing Systems*, 34:7054–7067, 2021.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Xiao Zhang, Jinghui Chen, Quanquan Gu, and David Evans. Understanding the intrinsic robustness of image distributions using conditional generative models. In *International conference on artificial intelligence and statistics*, pp. 3883–3893. PMLR, 2020.