Benign Overfitting in Adversarially Robust Linear Classification

Jinghui Chen[1*] **Yuan Cao**[2*] **Quanquan Gu**[3]

[1]The Pennsylvania State University, jzc5917@psu.edu
[2]The University of Hong Kong, yuancao@hku.hk
[3]University of California, Los Angeles, qgu@cs.ucla.edu
[*]Equal contribution

Abstract

"Benign overfitting", where classifiers memorize noisy training data yet still achieve a good generalization performance, has drawn great attention in the machine learning community. To explain this surprising phenomenon, a series of works have provided theoretical justification for over-parameterized linear regression, classification, and kernel methods. However, it is not clear if benign overfitting can occur in the presence of adversarial examples, i.e., examples with tiny and intentional perturbations to fool the classifiers. In this paper, we show that benign overfitting indeed occurs in adversarial training, a principled approach to defend against adversarial examples, on subGaussian mixture data. In detail, we prove the risk bounds of the adversarially trained linear classifier on the mixture of sub-Gaussian data under ℓ_p adversarial perturbations. Our result suggests that under moderate perturbations, adversarially trained linear classifiers can achieve the near-optimal standard and adversarial risks, despite overfitting the noisy training data. Numerical experiments validate our theoretical findings.

1 INTRODUCTION

Modern machine learning methods such as deep learning have made many breakthroughs in a variety of application domains, including image classification [He et al., 2016a, Krizhevsky et al., 2012], speech recognition [Hinton et al., 2012] and etc. These models are typically over-parameterized: the number of model parameters far exceeds the size of the training samples. One mystery is that, these over-parameterized models can memorize noisy training data and yet still achieve quite good generalization performances on the test data [Zhang et al., 2017]. Many efforts have been made to explain this striking phenomenon, which against what the classical notion of overfitting might suggest. A line of research works [Soudry et al., 2018, Ji and Telgarsky, 2019b, Nacson et al., 2019, Gunasekar et al., 2018b,a] shows that there exists the so-called implicit bias [Neyshabur, 2017]: the training algorithms tend to converge to certain kinds of solutions even with no explicit regularization. Specifically, Soudry et al. [2018], Ji and Telgarsky [2019b], Nacson et al. [2019] demonstrate that gradient descent trained linear classifiers on logistic or exponential loss with no regularization asymptotically converge to the maximum L_2 margin classifier. Recent works [Bartlett et al., 2020, Chatterji and Long, 2020, Cao et al., 2021, Wang and Thrampoulidis, 2021, Tsigler and Bartlett, 2020] further shows that over-parameterized and implicitly regularized interpolators can indeed achieve small test error, and formulate this phenomenon as "benign overfitting". More concretely, suppose the classification model f is parameterized by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and the loss is denoted as $\ell(\cdot)$. The population risk is define as

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[f_{\boldsymbol{\theta}}(\mathbf{x}) \neq y],$$

where data pair $(\mathbf{x}, y)$ is generated from certain data generation model. Chatterji and Long [2020] shows that with sufficient over-parameterization, gradient descent trained maximum L_2 margin classifier can achieve nearly optimal population risk on noisy data for data generated from a sub-Gaussian mixture model. This suggests that the overfitting can be "benign" in the over-parameterized setting.

Besides these studies on the benign overfitting phenomenon, another well-known feature of modern machine learning methods is that they are vulnerable to adversarial examples. Recent studies [Szegedy et al., 2013, Goodfellow et al., 2015] show that modern machine learning systems are brittle: slight input perturbation that is imperceptible to human eyes could mislead a well-trained classifier into wrong classification result. These malicious inputs are also known as the adversarial examples [Szegedy et al., 2013, Goodfellow et al., 2015]. Adversarial examples raise severe trustworthy issues and security concerns on the current machine learning systems especially in security-critical applications. Various methods [Kurakin et al., 2016, Madry et al., 2018, Zhang et al., 2019, Wang et al., 2019, 2020] have been proposed to defend against the threats posed by adversarial examples. One of the notable approaches is adversarial training [Madry et al., 2018]. Specifically, adversarial training solves the following min-max optimization problem,

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n}\sum_{i=1}^{n} \max_{\mathbf{x}_i'\in\mathcal{B}_\epsilon^p(\mathbf{x}_i)} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i'), y_i),$$

where $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is the training set and $\mathcal{B}_\epsilon^p(\mathbf{x}_i) = \{\mathbf{x} : \|\mathbf{x}-\mathbf{x}_i\|_p \le \epsilon\}$ denotes the ϵ-ball around $\mathbf{x}_i$ in ℓ_p norm ($p \ge 1$). Many empirical or theoretical studies have been conducted trying to analyze or further improve adversarial training robustness [Zhang et al., 2019, Wang et al., 2020, Carmon et al., 2019, Wang et al., 2019, Raghunathan et al., 2020]. A recent work [Sanyal et al., 2021] also pointed out that normally trained interpolators with the presence of label noise are unlikely to be adversarially robust, while adversarially robust classifiers cannot overfit noisy labels under certain conditions. Rice et al. [2020] showed that overfitting can hurt robust generalization in adversarial training on several real-world datasets. Dong et al. [2021] pointed out that robust overfitting in adversarial training is caused by memorizing one-hot labels, which can be relieved by proper regularization. However, there still lacks theoretical understanding why and when benign overfitting can or cannot occur for adversarial training.

In this paper, we show that benign overfitting can indeed occur in adversarial training under certain data distributions, which largely advances our understanding towards overfitting in adversarial training. We summarize our contributions of this paper in the following

- We show that the benign overfitting phenomenon can occur in adversarially robust linear classifiers with sufficient over-parameterization for data generated from a Gaussian mixture model. Specifically, under moderate ℓ_p norm perturbations, adversarially trained linear classifiers can achieve the near-optimal standard and adversarial risks, in spite of overfitting the noisy training data.
- When the perturbation strength ϵ is set to be 0, our adversarial risk bound reduces to the standard one. The resulting standard risk bound extends Chatterji and Long [2020]'s risk bound to further characterize the behavior of the linear classifier trained by t-step gradient descent.
- We show that depending on the value of p (perturbation norm), the adversarial risk bound can be different. The higher value of p (typically for $p \ge 2$ case) actually leads to a larger gap between the adversarial risk and the standard risk with the same ϵ.

Complementary to our theory, we also conduct numerical experiments to show that if certain data distribution assumption is violated, overfitting can become harmful.

Notation. we use lower case letters to denote scalars and lower case bold face letters to denote vectors. For a vector $\mathbf{x} \in \mathbb{R}^d$, we denote its ℓ_p norm ($p \ge 1$) of $\mathbf{x}$ by $\|\mathbf{x}\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$, the ℓ_∞ norm of $\mathbf{x}$ by $\|\mathbf{x}\|_\infty = \max_{i=1}^d |x_i|$. We denote $\mathbf{x}^{\circ p}$ as the element-wise p-power of $\mathbf{x}$. The notation $(\mathbf{x}, y) \sim \mathcal{D}$ denotes that the data pair $(\mathbf{x}, y)$ is generated from a distribution $\mathcal{D}$. For $p \ge 1$, we denote $\mathcal{B}_r^p(\mathbf{x})$ as the ℓ_p norm ball of radius r centered at $\mathbf{x}$. Given two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists a constant $0 < C < +\infty$ such that $a_n \le C\, b_n$. We denote $a_n = \Omega(b_n)$ if $b_n = O(a_n)$. We denote $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$.

2 RELATED WORK

There exists a large body of works on adversarial training, implicit bias and benign overfitting. In this section, we review the most relevant works to ours.

Adversarial Training. Adversarial training [Madry et al., 2018] and its variants [Zhang et al., 2019, Wang et al., 2019, 2020] are currently the most effective type of approaches to empirically defend against adversarial examples [Szegedy et al., 2013, Goodfellow et al., 2015]. Many attempts have been made to understand its empirical success. Charles et al. [2019], Li et al. [2020] showed that the adversarially trained linear classifier directionally converges to the maximum margin classifier. Gao et al. [2019], Zhang et al. [2020b] showed that adversarial training with neural networks can achieve low robust training loss. Yet these conclusions cannot explain the test (population) performances. Another line of research focuses on the generalization performance of adversarial training and the number of training samples. Schmidt et al. [2018] showed that adversarial models require more data than standard models to achieve certain test accuracy. Chen et al. [2020] showed that more data may actually increase the gap between the generalization error of adversarially-trained models and standard models. Yin et al. [2019], Cullina et al. [2018] studied the adversarial Rademacher complexity and VC-dimensions. Some other works focus on the trade-off between robustness and natural accuracy [Zhang et al., 2019, Tsipras et al., 2019, Wu et al., 2020, Raghunathan et al., 2020, Yang et al., 2020, Dobriban et al., 2020, Javanmard and Soltanolkotabi, 2020], adversarial model complexity lower bound [Allen-Zhu and Li, 2020], as well as the provable robustness upper bound [Fawzi et al., 2018, Zhang et al., 2020a]. Liu et al. [2021] studied the impact of hard training instances on adversarially trained model's overfitting behavior.

Recently, some works also focus on studying the learning of robust halfspaces and linear models. Montasser et al. [2020] studied the conditions on the adversarial perturbation sets under which halfspaces are robustly learnable in the presence of random label noise. Diakonikolas et al. [2020] studied the computational complexity of adversarially robust halfspaces under ℓ_p norm perturbations. Zou et al. [2021a] showed that adversarially trained halfspaces are provably robust with low robust classification error in the presence of noise. Dan et al. [2020] proposed an adversarial signal to noise ratio and studied the excess risk lower/upper bounds for learning Gaussian mixture models. Taheri et al. [2020], Javanmard and Soltanolkotabi [2020] studied adversarial learning of linear models on Gaussian mixture data, where the data dimension and the number of training data points have a fixed ratio.

Implicit Bias. Several recent works studied the implicit bias of various training algorithms in over-parameterized models. Soudry et al. [2018] studied the implicit bias of gradient descent trained on linearly separable data while Ji and Telgarsky [2019b] studied non-separable cases. Gunasekar et al. [2018a] studied the implicit bias of various optimization methods in linear regression and classification problems. Ji and Telgarsky [2019a] studied the implicit bias for deep linear networks and Arora et al. [2019], Gunasekar et al. [2018b] studied the implicit bias for matrix factorization. Lyu and Li [2020] studied implicit regularization of homogeneous neural networks with exponential loss and logistic loss.

Benign Overfitting and Double Descent. A series of recent works have studied the "benign overfitting" phenomenon Bartlett et al. [2020] that when training over-parameterized models, classifiers can still achieve good population risk even when overfitting the noisy training data. Bartlett et al. [2020], Tsigler and Bartlett [2020] studied the risk bounds for over-parameterized linear (ridge) regression and showed that under certain settings, the interpolating linear model with minimum parameter norm can have asymptotically optimal risk. Belkin et al. [2018, 2019a,b], Hastie et al. [2019], Wu and Xu [2020] quantified the dependency curve between the population risk and the degree of over-parameterization and showed that the curve has a double-descent shape. Chatterji and Long [2020] studied the risk bounds in over-parameterized linear logistic regression with label flipping noises. Cao et al. [2021] further tighten the risk bound in Chatterji and Long [2020] in low SNR settings. Zou et al. [2021b] studied benign overfitting of stochastic gradient descent for linear regression. Shamir [2022] studied benign overfitting for linear predictors using a generic data model. Recently, Frei et al. [2022], Cao et al. [2022] studied the benign overfitting in two-layer fully-connected neural networks or CNNs. Chatterji and Long [2022] showed a negative result for basis pursuit and compared it with ordinary least squares. Wald et al. [2022] showed some negative results and suggests that the phenomenon of benign overfitting might not favorably extend to settings in which robustness or fairness are desirable.

3 PROBLEM SETTING AND PRELIMINARIES

In order to properly characterize the benign overfitting phenomenon in adversarial training, we also define the population adversarial risk, which is the counterpart for population risk in the standard training scenario:

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\exists \mathbf{x}' \in \mathcal{B}_\epsilon^p(\mathbf{x}) \ s.t., \ f_{\boldsymbol{\theta}}(\mathbf{x}') \neq y\big].$$

The adversarial risk measures the misclassification rate of the target classifier under the presence of ℓ_p-norm adversarial perturbations. It is easy to observe that the adversarial risk is always larger than the standard risk as it requires the classifier to correctly classify the data examples within the entire local ℓ_p norm ball.

We consider a sub-Gaussian mixture data generation model in our work. Specifically, the clean data $(\tilde{\mathbf{x}}, \tilde{y}) \sim \tilde{\mathcal{D}}$ is generated such that, for each data point $(\tilde{\mathbf{x}}, \tilde{y}) \in \mathbb{R}^d \times \{\pm 1\}$, we have $\tilde{y} \sim \text{Unif}(\{\pm 1\})$ and $\tilde{\mathbf{x}} = \tilde{y}\boldsymbol{\mu} + \boldsymbol{\xi}$ where $\boldsymbol{\xi} \in \mathbb{R}^d$ and $\xi_1, \xi_2, \ldots, \xi_d$ are i.i.d. zero-mean sub-Gaussian variables with sub-Gaussian norm at most 1. The actual data examples are sampled from a noisy distribution $\mathcal{D}$ which is close to the clean distribution $\tilde{\mathcal{D}}$. Specifically, $\mathcal{D}$ can be any distribution over $\mathbb{R}^d \times \{\pm 1\}$ who has the same marginal distribution on $\mathbb{R}^d$ and the total variation distance $d_{\text{TV}}(\mathcal{D}, \tilde{\mathcal{D}}) \leq \eta$ where η denotes the noise level.

Note that our data generation model is standard for studying the population risk of over-parameterized linear classification. In fact, it is exactly the same as the one studied in Chatterji and Long [2020]. In this model, following standard coupling lemma [Lindvall, 2002], there always exists a joint distribution on original data and noisy data $((\tilde{\mathbf{x}}, \tilde{y}), (\mathbf{x}, y))$ such that the marginal distribution for $(\tilde{\mathbf{x}}, \tilde{y})$ is $\tilde{\mathcal{D}}$, the marginal distribution for $(\mathbf{x}, y)$ is $\mathcal{D}$, $\mathbb{P}[\mathbf{x} = \tilde{\mathbf{x}}] = 1$ and $\mathbb{P}[y \neq \tilde{y}] \leq \eta$.

In this paper, we study the problem of robust binary classification with training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ drawn i.i.d. from the distribution $\mathcal{D}$. Let's denote the "clean" sample index as $\mathcal{C} := \{k : y_k = \tilde{y}_k\}$ and the "noisy" sample index as $\mathcal{N} := \{k : y_k \neq \tilde{y}_k\}$. We consider the adversarially trained linear classifier under exponential loss. In such cases, the adversarial loss can be explicitly written as

$$L(\boldsymbol{\theta}) = \sum_{i=1}^n \max_{\mathbf{x}_i' \in \mathcal{B}_\epsilon^p(\mathbf{x}_i)} \exp(-y_i \boldsymbol{\theta}^\top \mathbf{x}_i'). \qquad (1)$$

In gradient descent adversarial training algorithm, the adversarial loss $L(\boldsymbol{\theta})$ is minimized by first solving the inner maximization problem in (1) with respect to the current model parameter $\boldsymbol{\theta}_{t-1}$ and then update the model parameter $\boldsymbol{\theta}_t$ by performing gradient descent to minimize the adversarial loss in each iteration. We summarized the training procedure for gradient descent adversarial training[1] in Algo-

[1] Note that in practice people often initialize $\boldsymbol{\theta}_0$ by a small ran-

Algorithm 1 Gradient Descent Adversarial Training

```
input: Training data $\{\mathbf{x}_i, y_i\}_{i=1}^n$, number of training iterations $T$, maximum perturbation strength $\epsilon$, training step sizes $\alpha_t$;
initialize model parameter $\boldsymbol{\theta}_0 = \mathbf{0}$
for $t = 1, \ldots, T$ do
    for each $\{\mathbf{x}_i, y_i\}$ do
        $\mathbf{x}_i' = \mathrm{argmax}_{\mathbf{x}_i' \in \mathcal{B}_\epsilon^p(\mathbf{x}_i)} \exp(-y_i \boldsymbol{\theta}_{t-1}^\top \mathbf{x}_i')$
    end for
    $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha_t \cdot \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_{t-1})$
end for
```

rithm 1. Note that in the linear classifier setting, the inner maximization problem in (1) has the following property

$$
\begin{aligned}
\underset{\mathbf{x}_i' \in \mathcal{B}_\epsilon^p(\mathbf{x}_i)}{\mathrm{argmax}} \exp(-y_i \boldsymbol{\theta}^\top \mathbf{x}_i') &= \underset{\mathbf{u}_i \in \mathcal{B}_\epsilon^p(\mathbf{0})}{\mathrm{argmax}} \exp(-y_i \boldsymbol{\theta}^\top (\mathbf{x}_i + \mathbf{u}_i)) \\
&= \underset{\|\mathbf{u}_i\|_p \le \epsilon}{\mathrm{argmin}}\, y_i \boldsymbol{\theta}^\top \mathbf{u}_i. \qquad (2)
\end{aligned}
$$

By Hölders' inequality, it is easy to observe that the optimal adversarial loss and the corresponding gradient can be written as

$$
\begin{aligned}
&L(\boldsymbol{\theta}) = \sum_{i=1}^n \exp(-y_i \boldsymbol{\theta}^\top \mathbf{x}_i + \epsilon \|\boldsymbol{\theta}\|_q), \\
&\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \\
&= -\sum_{i=1}^n (y_i \mathbf{x}_i - \epsilon \cdot \partial \|\boldsymbol{\theta}\|_q) \exp(-y_i \boldsymbol{\theta}^\top \mathbf{x}_i + \epsilon \|\boldsymbol{\theta}\|_q),
\end{aligned}
$$

where $1/p + 1/q = 1$. Also note that in the over-parameterized settings, the training examples drawn from our data generation model are linearly separable with high probability (See Lemma 12 in Section 5). Linearly separable property ensures that the training samples have a positive margin (with high probability). Following Li et al. [2020], we also define the standard and adversarial margin as

$$
\begin{aligned}
\bar{\gamma} &:= \max_{\|\boldsymbol{\theta}\|_q = 1} \min_{i \in [n]} y_i \boldsymbol{\theta}^\top \mathbf{x}_i, \\
\gamma &:= \max_{\|\boldsymbol{\theta}\|_2 = 1} \min_{i \in [n]} \min_{\mathbf{x}_i' \in \mathcal{B}_\epsilon^p(\mathbf{x}_i)} y_i \boldsymbol{\theta}^\top \mathbf{x}_i', \qquad (3)
\end{aligned}
$$

which are useful in our later analysis. We also define the unique linear classifier θ that achieves adversarial margin γ defined above as $\mathbf{w}$.

4 MAIN RESULTS

In this section, we study both the behavior of the population risk and the population adversarial risk for adversarially trained linear classifiers.

Assumption 1. *The adversarial perturbation radius ϵ is upper bounded by a constant R and is smaller than the ℓ_p data margin $\bar{\gamma}$, i.e., $\epsilon \le \min\{R, \bar{\gamma}\}$.*

The goal of adversarial training is to obtain high-accuracy classifiers that are also robust to small input perturbations which can be ignored by human beings (e.g., small ℓ_∞-norm perturbations that are invisible to human eyes). Therefore, Assumption 1 is reasonable by constraining the maximum allowable perturbation magnitude.

Assumption 2. *The noise $\boldsymbol{\xi}$ in the data generation model satisfies that $\mathbb{E}[\|\boldsymbol{\xi}\|_2^2] \ge \kappa d$ for some constant κ.*

Assumption 2 is a common condition that has also been considered in Chatterji and Long [2020]. It ensures that the summation of the variances of the data input increases in the order of $\Theta(d)$. Clearly, this assumption covers the most common setting where the entries of ξ are i.i.d. and have a variance larger than or equal to κ.

Assumption 3. *The gradient descent starts at* $\mathbf{0}$, *and the step sizes are set as $\alpha_0 = 1/(Gdn)$, $\alpha_t = \alpha \le 1/(GdnM)$ for $M = \max\{[2d + \epsilon(q-1)d^{\frac{3q-2}{2q-2}}/\gamma] \exp(-\gamma^2/(Gd) + \epsilon/G), 1\}$ and a constant G.*

Assumption 3 summarizes our assumptions about the gradient descent algorithm on the adversarial loss. The learning rate conditions here are to ensure the convergence of adversarial training, and is inspired by Li et al. [2020].

We first present our theorem for standard risk of adversarial training method (Algorithm 1).

Theorem 4 (Standard Risk of Adversarial Training)**.** *For any $p \in [1, +\infty)$, suppose that Assumptions 1, 2 and 3 hold with $\kappa \in (0, 1]$ and large enough constants R and G. Moreover, for any $\delta \in (0, 1)$, suppose the number of training samples $n \ge C \log(1/\delta)$, the dimension $d \ge C \cdot \max\{n\|\boldsymbol{\mu}\|_2^2, n^2 \log(n/\delta)\}$, the noise level $\eta < 1/C$, and $\|\boldsymbol{\mu}\|_2^2 \ge C \max\{\log(n/\delta), \epsilon\|\boldsymbol{\mu}\|_q\}$ for a large enough constant C. Then with probability at least $1 - \delta$, adversarially trained linear classifier $f_{\boldsymbol{\theta}_t}$ for sufficiently large t under ℓ_p-norm ϵ-perturbation satisfies the following standard risk*

$$
\begin{aligned}
&\mathbb{P}_{(\mathbf{x}, y) \sim \mathcal{D}}[f_{\boldsymbol{\theta}_t}(\mathbf{x}) \ne y] \\
&\le \eta + \exp\left(- C' \left(\frac{(\|\boldsymbol{\mu}\|_2^2 - 4\epsilon \|\boldsymbol{\mu}\|_q)}{(C'' + \epsilon)\sqrt{d}} - \frac{C''' \|\boldsymbol{\mu}\|_2 \log n}{\log t} \right)^2 \right),
\end{aligned}
$$

where $C', C'', C''' > 0$ are absolute constants, $1/p + 1/q = 1$.

Remark 5. *Theorem 4 presents the standard risk of adversarial training under ℓ_p norm perturbations. Note that adversarially trained linear classifier enjoys a bounded population risk which decreases as the number of training*

dom vector (e.g., Xavier initialization [Glorot and Bengio, 2010]), while we follow Li et al. [2020] and set $\boldsymbol{\theta}_0 = \mathbf{0}$ for the ease of theoretical analysis.

iterations t increases. Specifically, when $t \to \infty$, we have

$$\lim_{t\to\infty} \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[f_{\boldsymbol{\theta}_t}(\mathbf{x}) \neq y] \\ \leq \eta + \exp\Bigg(- C'\Big(\frac{\big(\|\boldsymbol{\mu}\|_2^2 - 4\epsilon\|\boldsymbol{\mu}\|_q\big)}{(C''+\epsilon)\sqrt{d}}\Big)^2\Bigg). \quad (4)$$

Remark 6. *For (4), consider the case when the sample size n is fixed but dimension d and $\|\boldsymbol{\mu}\|_2$ are growing, we discuss the conditions to reach minimum standard risk of noise level η. Note that when $1 \leq p \leq 2$ we have $q \geq 2$ and $\|\boldsymbol{\mu}\|_q \leq \|\boldsymbol{\mu}\|_2$. In this case, if $\|\boldsymbol{\mu}\|_2 = \Omega(d^{1/4})$, the standard risk will come close to the noise level η when d is sufficiently large. When $p > 2$ and therefore $q < 2$, we have $\|\boldsymbol{\mu}\|_q \leq d^{1/q-1/2}\|\boldsymbol{\mu}\|_2$. In this case, if $\|\boldsymbol{\mu}\|_2 = \Omega(d^{1/4})$ and $\epsilon = O(\|\boldsymbol{\mu}\|_2/d^{1/q-1/2})$, the standard risk will come close to the noise level η with sufficiently large d. Note that our theorem condition also requires that $\|\boldsymbol{\mu}\|_2 = O(\sqrt{d})$. Therefore, in order to reach the standard risk of η, we need $\|\boldsymbol{\mu}\|_2 = \Theta(d^r)$ for some $r \in (1/4, 1/2]$.*

Remark 7. *Choosing $\epsilon = 0$ will reduce to the standard training case. Specifically, if we set $\epsilon = 0$ in (4), it reduces to the same conclusion as Theorem 3.1 in Chatterji and Long [2020]. However, our result is more general, as it covers the setting of adversarial training and gives risk bounds for the linear model obtained with a finite number of gradient descent iterations.*

Theorem 8 (Adversarial Risk of Adversarial Training)**.** *For any $\delta \in (0,1)$, under the same conditions as in Theorem 4, with probability at least $1-\delta$, the adversarially trained linear classifier $f_{\boldsymbol{\theta}_t}$ for sufficiently large t under ℓ_p-norm ϵ-perturbation satisfies the following adversarial risk if $1 \leq p \leq 2$*

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\exists \mathbf{x}' \in \mathcal{B}_\epsilon^p(\mathbf{x}) \; s.t., \; f_{\boldsymbol{\theta}}(\mathbf{x}') \neq y\big] \\ \leq \eta + \exp\Bigg(- C'\Big(\frac{\big(\|\boldsymbol{\mu}\|_2^2 - 4\epsilon\|\boldsymbol{\mu}\|_q\big)}{(C''+\epsilon)\sqrt{d}} - \frac{C'''\|\boldsymbol{\mu}\|_2 \log n}{\log t} - \epsilon\Big)^2\Bigg),$$

and if $p > 2$,

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\exists \mathbf{x}' \in \mathcal{B}_\epsilon^p(\mathbf{x}) \; s.t., \; f_{\boldsymbol{\theta}}(\mathbf{x}') \neq y\big] \\ \leq \eta + \exp\Bigg(- C'\Big(\frac{\big(\|\boldsymbol{\mu}\|_2^2 - 4\epsilon\|\boldsymbol{\mu}\|_q\big)}{(C''+\epsilon)\sqrt{d}} - \frac{C'''\|\boldsymbol{\mu}\|_2 \log n}{\log t} - \epsilon d^{\frac{1}{q}-\frac{1}{2}}\Big)^2\Bigg),$$

where $C', C'', C''' > 0$ are absolute constants, $1/p+1/q = 1$.

Remark 9. *Theorem 8 shows the adversarial risk of adversarial training under ℓ_p norm perturbations. The major difference from the standard risk (Theorem 4) lies in the additional ϵ or $\epsilon d^{1/q-1/2}$ term in the exponential function. This aligns with the common sense that adversarial risk should always be larger than the standard risk. This also suggests that for larger p-norm ($p > 2$) perturbation, the same magnitude of perturbation would lead to a larger gap between the adversarial risk and the standard risk. In terms of the perturbation strength, we can also observe that with a larger ϵ, adversarially trained classifiers obtain worse adversarial risk. This has been verified by many empirical observations of adversarial training [Madry et al., 2018, Zhang et al., 2019].*

Remark 10. *Note that when $t \to \infty$, if $1 \leq p \leq 2$, we have the following adversarial risk bound:*

$$\lim_{t\to\infty} \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\exists \mathbf{x}' \in \mathcal{B}_\epsilon^p(\mathbf{x}), f_{\boldsymbol{\theta}}(\mathbf{x}') \neq y\big] \\ \leq \eta + \exp\Bigg(- C'\Big(\frac{\big(\|\boldsymbol{\mu}\|_2^2 - 4\epsilon\|\boldsymbol{\mu}\|_q\big)}{(C''+\epsilon)\sqrt{d}} - \epsilon\Big)^2\Bigg),$$

and if $p > 2$, we have

$$\lim_{t\to\infty} \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\exists \mathbf{x}' \in \mathcal{B}_\epsilon^p(\mathbf{x}), f_{\boldsymbol{\theta}}(\mathbf{x}') \neq y\big] \\ \leq \eta + \exp\Bigg(- C'\Big(\frac{\big(\|\boldsymbol{\mu}\|_2^2 - 4\epsilon\|\boldsymbol{\mu}\|_q\big)}{(C''+\epsilon)\sqrt{d}} - \epsilon d^{\frac{1}{q}-\frac{1}{2}}\Big)^2\Bigg).$$

Similar to the standard risk case (Remark 6), when $1 \leq p \leq 2$, if $\|\boldsymbol{\mu}\|_2 = \Theta(d^r)$ for some $r \in (1/4, 1/2]$, the adversarial risk will also come close to the noise level η with sufficiently large d. When $p > 2$, if we have $\|\boldsymbol{\mu}\|_2 = \Theta(d^r)$ for some $r \in (1/4, 1/2]$ and $\epsilon = O(\|\boldsymbol{\mu}\|_2/d^{1/q})$, the adversarial risk will be close to η with sufficiently large d. Note that compared to the standard risk, this requirement on ϵ is slightly stronger.

Remark 11. *Note that our results in Theorem 8 imply that overfitting in adversarial training can be benign for certain distributions (e.g., subGaussian mixture data). This is later empirically verified in the experiments for both linear and neural network models.*

5 PROOF OUTLINE OF THE MAIN RESULTS

In this section, we present the proof of our main theorems, which consists of three main steps.

Statistical properties of the training data points. We first list some basic properties of the training data points based on our data model defined in Section 3.

Lemma 12 (Lemma 4.7 in Chatterji and Long [2020])**.** *Let $\mathbf{z}_k = y_k\mathbf{x}_k$. There exist absolute constants R, κ and G and C, such that if the assumptions in Theorem 4 hold, then with probability at least $1-\delta$,*

$$\frac{d}{c_0} \leq \|\mathbf{z}_k\|_2^2 \leq c_0 d \textit{ for all } k \in [n], \quad (5)$$

$$|\mathbf{z}_i^\top \mathbf{z}_j| \leq c_0\big(\|\boldsymbol{\mu}\|_2^2 + \sqrt{d\log(n/\delta)}\big) \textit{ for all } i \neq j, \quad (6)$$

$$|\boldsymbol{\mu}^\top \mathbf{z}_k - \|\boldsymbol{\mu}\|_2^2| \leq \|\boldsymbol{\mu}\|_2^2/2 \textit{ for all } k \in \mathcal{C}, \quad (7)$$

$$|\boldsymbol{\mu}^\top \mathbf{z}_k - (-\|\boldsymbol{\mu}\|_2^2)| \leq \|\boldsymbol{\mu}\|_2^2/2 \textit{ for all } k \in \mathcal{N}, \quad (8)$$

the number of noisy samples $|\mathcal{N}| \leq (\eta + c_1)n$, and all training samples are linearly separable, where $c_0 > 1$ is an absolute constant.

Lemma 12 directly follows Lemma 4.7 in Chatterji and Long [2020]. It provides direct high probability bounds for $\|\mathbf{z}_k\|_2$ and $\boldsymbol{\mu}^\top \mathbf{z}_k$ and also suggests that $\mathbf{z}_k$ vectors are nearly pairwise orthogonal in over-parameterized settings. It also guarantees that training examples are linearly separable with high probability.

Landscape properties of the training objective function. Given the properties of the training data points, we proceed to establish the landscape properties of the objective function $L(\boldsymbol{\theta}_1)$. The following lemma bound the loss for the adversarially trained classifier in step 1.

Lemma 13. *[Theorem 3.4 in Li et al. [2020]] Under the same conditions as in Theorem 4, with probability at least $1-\delta$, we have $L(\boldsymbol{\theta}_1) \le 2n$, and*

$$L(\boldsymbol{\theta}_{t+1}) \le L(\boldsymbol{\theta}_t), \tag{9}$$

$$1 - \frac{\boldsymbol{\theta}_t^\top \mathbf{w}}{\|\boldsymbol{\theta}_t\|_2} \le \frac{c_3 \log n}{\log t} \tag{10}$$

for all $t > 0$, where c_3 is an absolute constant.

By Lemma 13, one can easily observe that the adversarial training loss is bounded by $2n$ along the entire training trajectory. Lemma 13 also suggests that when $t \to \infty$, the adversarially trained classifier $\boldsymbol{\theta}_t$ will converge in direction to the max adversarial margin classifier $\mathbf{w}$ defined in (3).

Length and direction of the adversarial training iterates $\boldsymbol{\theta}_t$. We also establish the properties of the adversarial training iterates $\boldsymbol{\theta}_t$. We have the following lemmas.

Lemma 14. *Under the same conditions as in Theorem 4, for all adversarial training iteration $t > 0$, with probability at least $1-\delta$, we have $\|\boldsymbol{\theta}_{t+1}\|_2 \le (\sqrt{c_0} + \epsilon)\sqrt{d}\sum_{m=0}^{t} \alpha_m L(\boldsymbol{\theta}_m)$, where c_0 is the absolute constant in Lemma 12.*

Lemma 14 upper bound the L_2 norm of adversarially trained classifier $\boldsymbol{\theta}_t$ by the summation of training losses along the training trajectory.

Lemma 15. *Let $\mathbf{z}_k = y_k \mathbf{x}_k$, under the same conditions as in Theorem 4, for all adversarial training iteration $t \ge 0$, with probability as least $1-\delta$, we have $\max_{k=1}^n \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_k) \le c_3 \min_{k=1}^n \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_k)$, where $c_3 > 0$ is an absolute constant.*

Lemma 15 provides us a way to control the loss the noisy examples during the training procedure. Note that if $\max_{k=1}^n \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_k) \le c_3 \min_{k=1}^n \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_k)$, we also have $\max_{k=1}^n \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_k + \epsilon\|\boldsymbol{\theta}_t\|_q) \le c_3 \min_{k=1}^n \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_k + \epsilon\|\boldsymbol{\theta}_t\|_q)$. Therefore, the worst example training loss can be bounded via the best example training loss and further be bounded by the average training loss $L(\boldsymbol{\theta}_t)$. In this way, we can guarantee that those noisy examples will not have major influence on model training even in later training stages.

By Lemmas 12-15, we establish the following key lemma for our main theorems.

Lemma 16. *Under the same condition as in Theorem 4, with probability at least $1-\delta$, the adversarially trained linear model parameter $\boldsymbol{\theta}_t$ satisfies*

$$\frac{\boldsymbol{\mu}^\top \boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|_2} \ge \left(\frac{\|\boldsymbol{\mu}\|_2^2}{4} - \epsilon\|\boldsymbol{\mu}\|_q\right)\frac{1}{(\sqrt{c_0}+\epsilon)\sqrt{d}} - \frac{c_3\|\boldsymbol{\mu}\|_2 \log n}{\log t}.$$

where c_0 is the absolute constant in Lemma 12.

Lemma 16 provides the lower bound for the inner product of $\boldsymbol{\mu}$ and the direction of $\boldsymbol{\theta}_t$. This lemma extends Lemma 4.4 in Li et al. [2020] by considering the training iteration t rather than just the converged classifier $\mathbf{w}$, and also extends to the adversarial training setting. Notice that this lower bound actually gets larger with the increase of iteration t.

Finalizing the proof. We now present the proof for Theorems 4 and 8.

Proof of Theorem 4. First, following standard coupling lemma [Lindvall, 2002], there always exists a joint distribution on original data and noisy data $((\tilde{\mathbf{x}}, \tilde{y}), (\mathbf{x}, y))$ such that the marginal distribution for $(\tilde{\mathbf{x}}, \tilde{y})$ is $\tilde{\mathcal{D}}$, the marginal distribution for $(\mathbf{x}, y)$ is $\mathcal{D}$, $\mathbb{P}[\mathbf{x} = \tilde{\mathbf{x}}] = 1$ and $\mathbb{P}[y \ne \tilde{y}] \le \eta$. Notice that the standard population risk can be written as

$$\begin{aligned}\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[f_{\boldsymbol{\theta}_t}(\mathbf{x}) \ne y] &= \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y \cdot \boldsymbol{\theta}_t^\top \mathbf{x} < 0] \\ &\le \eta + \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[y \cdot \boldsymbol{\theta}_t^\top \mathbf{x} < 0, y = \tilde{y}] \\ &= \eta + \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[\tilde{y} \cdot \boldsymbol{\theta}_t^\top \mathbf{x} < 0],\end{aligned} \tag{11}$$

where the inequality holds since $\mathbb{P}[y \ne \tilde{y}] \le \eta$. Since $\tilde{y}$ is the clean label for $\mathbf{x}$, $\tilde{y}\mathbf{x}$ follows the same distribution as $\boldsymbol{\xi} + \boldsymbol{\mu}$ and $\mathbb{E}[\tilde{y} \cdot \hat{\boldsymbol{\theta}}^\top \mathbf{x}] = \hat{\boldsymbol{\theta}}^\top \boldsymbol{\mu}$. Therefore, (11) can be further written as

$$\begin{aligned}&\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[f_{\boldsymbol{\theta}_t}(\mathbf{x}) \ne y] \\ &\le \eta + \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\tilde{y} \cdot \boldsymbol{\theta}_t^\top \mathbf{x} - \mathbb{E}[\tilde{y} \cdot \boldsymbol{\theta}_t^\top \mathbf{x}] < -\boldsymbol{\theta}_t^\top \boldsymbol{\mu}\big] \\ &= \eta + \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\boldsymbol{\theta}_t^\top\big(\tilde{y}\mathbf{x} - \mathbb{E}[\tilde{y}\mathbf{x}]\big) < -\boldsymbol{\theta}_t^\top \boldsymbol{\mu}\big] \\ &\le \eta + \exp\left(-c\frac{(\boldsymbol{\theta}_t^\top \boldsymbol{\mu})^2}{\|\boldsymbol{\theta}_t\|_2^2}\right),\end{aligned} \tag{12}$$

where the last inequality holds by applying a Hoeffding-type concentration inequality (Theorem **??**) with $t = (\boldsymbol{\theta}_t^\top \boldsymbol{\mu})^2$. This bound in (12) enables the application of Lemma 16 which characterizes how the direction of $\boldsymbol{\theta}_t$ aligns with $\boldsymbol{\mu}$ during training. By direct calculation, we have

$$\begin{aligned}&\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[f_{\boldsymbol{\theta}_t}(\mathbf{x}) \ne y] \\ &\le \eta + \exp\left(-c\left(\frac{\left(\frac{\|\boldsymbol{\mu}\|_2^2}{4} - \epsilon\|\boldsymbol{\mu}\|_q\right)}{(\sqrt{c_0}+\epsilon)\sqrt{d}} - \frac{c_3\|\boldsymbol{\mu}\|_2 \log n}{\log t}\right)^2\right).\end{aligned}$$

This concludes the proof. □

Proof of Theorem 8. Similar to the proof of Theorem 4, we start with a calculating an upper bound of the population risk based on the formulation of the label noise. By the definition of the adversarial risk, we have

$$\begin{aligned}
&\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\exists\mathbf{x}' \in \mathcal{B}_\epsilon^p(\mathbf{x})\ s.t.,\ f_{\boldsymbol{\theta}_t}(\mathbf{x}') \neq y\big] \\
&= \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[\exists\mathbf{x}' \in \mathcal{B}_\epsilon^p(\mathbf{x})\ s.t.,\ y\cdot\boldsymbol{\theta}_t^\top\mathbf{x}' < 0] \\
&\leq \eta + \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[\exists\mathbf{x}' \in \mathcal{B}_\epsilon^p(\mathbf{x})\ s.t.,\ y\cdot\boldsymbol{\theta}_t^\top\mathbf{x}' < 0, y = \tilde{y}] \\
&= \eta + \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\Big[\min_{\mathbf{u}\in\mathcal{B}_\epsilon^p(\mathbf{0})} \tilde{y}\cdot\boldsymbol{\theta}_t^\top(\mathbf{x}+\mathbf{u}) < 0\Big] \\
&= \eta + \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\Big[\tilde{y}\cdot\boldsymbol{\theta}_t^\top\mathbf{x} - \epsilon\|\boldsymbol{\theta}_t\|_q < 0\Big], \qquad (13)
\end{aligned}$$

where the inequality holds in the same way as in (11). Since $\tilde{y}$ is the clean label for $\mathbf{x}$, $\tilde{y}\mathbf{x}$ follows the same distribution as $\boldsymbol{\xi}+\boldsymbol{\mu}$ and $\mathbb{E}[\tilde{y}\cdot\boldsymbol{\theta}_t^\top\mathbf{x}] = \boldsymbol{\theta}_t^\top\boldsymbol{\mu}$. Therefore, (13) can be further written as

$$\begin{aligned}
&\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\exists\mathbf{x}' \in \mathcal{B}_\epsilon^p(\mathbf{x})\ s.t.,\ f_{\boldsymbol{\theta}_t}(\mathbf{x}') \neq y\big] \\
&\leq \eta + \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\tilde{y}\cdot\boldsymbol{\theta}_t^\top\mathbf{x} - \mathbb{E}[\tilde{y}\cdot\boldsymbol{\theta}_t^\top\mathbf{x}] < -\boldsymbol{\theta}_t^\top\boldsymbol{\mu} + \epsilon\|\boldsymbol{\theta}_t\|_q\big] \\
&= \eta + \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\boldsymbol{\theta}_t^\top\big(\tilde{y}\mathbf{x} - \mathbb{E}[\tilde{y}\mathbf{x}]\big) < -\boldsymbol{\theta}_t^\top\boldsymbol{\mu} + \epsilon\|\boldsymbol{\theta}_t\|_q\big] \\
&\leq \eta + \exp\bigg(-c\frac{(\boldsymbol{\theta}_t^\top\boldsymbol{\mu} - \epsilon\|\boldsymbol{\theta}_t\|_q)^2}{\|\boldsymbol{\theta}_t\|_2^2}\bigg), \qquad (14)
\end{aligned}$$

where the second inequality holds by applying the Hoeffding-type concentration inequality (Theorem **??**) with $t = (\boldsymbol{\theta}_t^\top\boldsymbol{\mu} - \epsilon\|\boldsymbol{\theta}_t\|_q)^2$. Based on (14) and Lemma 16, we can further give the bounds of the adversarial risk. We consider the two settings $1 \leq p \leq 2$ and $2 < p < \infty$ separately.

When $1 \leq p \leq 2$, we have $q \geq 2$ and $\|\boldsymbol{\theta}\|_q \leq \|\boldsymbol{\theta}\|_2$. In this case, by Lemma 16 we obtain

$$\begin{aligned}
&\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[f_{\boldsymbol{\theta}_t}(\mathbf{x}) \neq y] \\
&\leq \eta + \exp\Bigg(-c\bigg(\frac{\big(\frac{\|\boldsymbol{\mu}\|_2^2}{4} - \epsilon\|\boldsymbol{\mu}\|_q\big)}{(\sqrt{c_0}+\epsilon)\sqrt{d}} - \frac{c_3\|\boldsymbol{\mu}\|_2\log n}{\log t} - \epsilon\bigg)^2\Bigg).
\end{aligned}$$

When $p > 2$ and therefore $q < 2$, we have $\|\boldsymbol{\mu}\|_q \leq d^{1/q-1/2}\|\boldsymbol{\mu}\|_2$. In this case, by Lemma 16 we obtain

$$\begin{aligned}
&\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[f_{\boldsymbol{\theta}_t}(\mathbf{x}) \neq y] \\
&\leq \eta + \exp\Bigg(-c\bigg(\frac{\big(\frac{\|\boldsymbol{\mu}\|_2^2}{4} - \epsilon\|\boldsymbol{\mu}\|_q\big)}{(\sqrt{c_0}+\epsilon)\sqrt{d}} - \frac{c_3\|\boldsymbol{\mu}\|_2\log n}{\log t} - \epsilon d^{\frac{1}{q}-\frac{1}{2}}\bigg)^2\Bigg).
\end{aligned}$$

This concludes the proof. □

6 EXPERIMENTS

In this section, we experimentally study the behavior of adversarially trained classifiers in the over-parameterized regime on both synthetic and real data.

6.1 SYNTHETIC DATA EXPERIMENTS

We generate 50 training samples and 2000 test samples and set the label noise ratio $\eta = 0.1$ for all experiments. Each clean sample $(\tilde{\mathbf{x}}, \tilde{y})$ is drawn from a Gaussian mixture model such that $\tilde{y} \sim \mathrm{Unif}(\{\pm 1\})$ and $\tilde{\mathbf{x}} = \tilde{y}\boldsymbol{\mu} + \boldsymbol{\xi}$ where $\boldsymbol{\xi} \in \mathbb{R}^d$ and $\xi_1, \xi_2, \ldots, \xi_d$ are i.i.d. standard Gaussian variables and $\boldsymbol{\mu}$ simply shares the same direction as an all-one vector but has various different magnitudes. This aligns with our model assumptions in Section 3. For the adversarial training algorithm, we directly follows Algorithm 1 except using a more practical Xavier normal initialization [Glorot and Bengio, 2010], i.e., sampling $\boldsymbol{\theta}_0$ i.i.d. from from $\mathcal{N}(0, 1/\sqrt{d})$. We set the learning rate $\alpha_t = 0.001$ and the total number of iterations $T = 1000$ for all experiments. All results are obtained by averaging over 10 independent runs (both data sampling and training).

In the first set of experiments, we verify our main conclusions in this paper that benign overfitting can occur in adversarial training. Figure 1 illustrates the risk and adversarial risk of adversarially trained linear classifiers versus the dimension d under different scalings of $\boldsymbol{\mu}$ for both ℓ_2-norm and ℓ_∞-norm perturbations. We can observe that when $\|\boldsymbol{\mu}\|_2 = d^{0.2}$, the (adversarial) risk starts to increase as the dimension d increases after an initial dive for both ℓ_2-norm and ℓ_∞-norm perturbations. While for cases where $\|\boldsymbol{\mu}\|_2 = d^{0.3}$ and $\|\boldsymbol{\mu}\|_2 = d^{0.4}$, we can observe that the (adversarial) risk decreases steadily to the optimal risk η as the dimension d increases. This result backup our theory in Section 4 that the optimal risk is achievable when $\|\boldsymbol{\mu}\|_2 = \Theta(d^r)$ and $r \in (1/4, 1/2]$. Note that the training error reaches 0 for all settings in Figure 1.

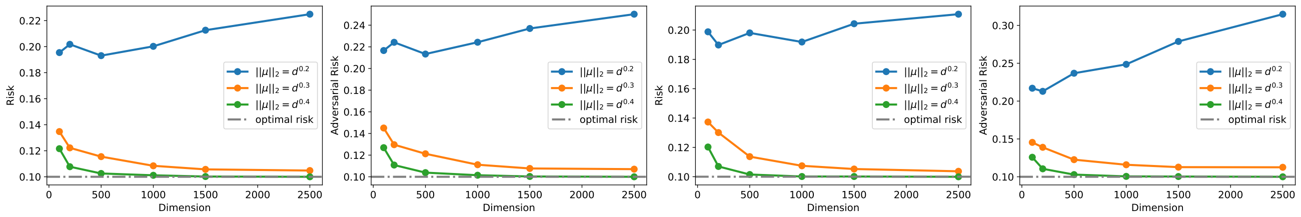
In Figure 2, we present the adversarial risk[2] of adversarially trained linear classifiers versus the training iterations t with different ϵ but fixed dimension d and $\|\boldsymbol{\mu}\|_2$ for both ℓ_2-norm and ℓ_∞-norm perturbations. We can also observe that in general, a larger ϵ will lead to the worse adversarial risk of the adversarially trained classifier. This also backs up our theory in Theorem 8.

As our ultimate goal is to study the benign overfitting phenomenon in real-world adversarial training settings, we also conducted experiments on 2-layer neural networks with ReLU activation functions. In fact, the performances on the 2-layer ReLU network suggest very similar trends as the linear model. Due to the space limit, we display these results in the supplemental materials.

6.2 REAL-WORLD DATA VERIFICATION

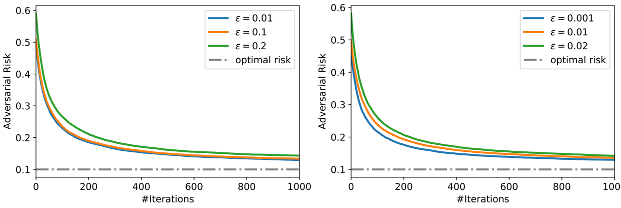
Rice et al. [2020] showed that overfitting in adversarial training can lead to worse empirical robustness on empirical

[2] Here we omit the plot for standard risk as the curves are essentially overlapping to each other.



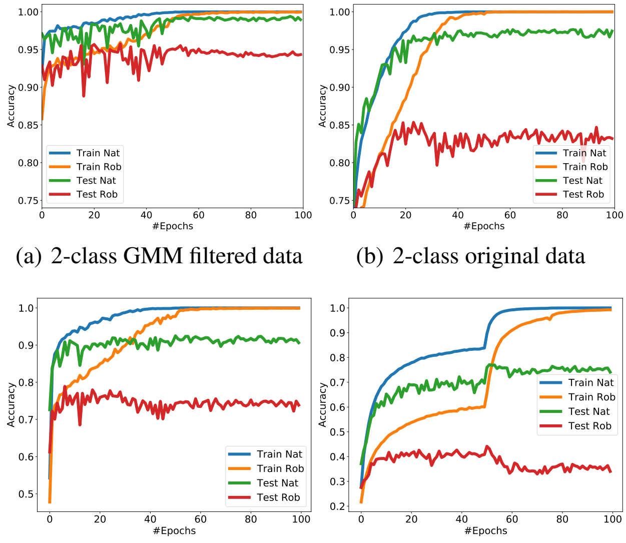
(a) ℓ_2 perturbation (b) ℓ_2 perturbation (c) ℓ_∞ perturbation (d) ℓ_∞ perturbation

Figure 1: Risk and adversarial risk of adversarially trained linear classifiers versus the dimension d under different scalings of $\boldsymbol{\mu}$. (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.



(a) ℓ_2 perturbation (b) ℓ_∞ perturbation

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



(a) 2-class GMM filtered data (b) 2-class original data

(c) 10-class GMM filtered data (d) 10-class original data

Figure 3: The learning curves for adversarial training [Madry et al., 2018] on CIFAR-10 data using GMM filtered data and the original data (a)(b) show the results for 2-class classification (airplane vs automobile) and (c)(d) show the results for 10-class classification.

image distributions such as CIFAR-10 [Krizhevsky et al., 2009] data. We want to ensure that our result is not contradict with their results since they are testing on empirical image distributions while our analysis is based on subGaussian mixture data, which CIFAR-10 data does not satisfy.

We conduct experiments to show that even for CIFAR-10 data, overfitting effect can be much less severe (or even benign) on robust classifiers, when we first filtered the input data by a Gaussian mixture model (GMM). Specifically, we craft a new dataset by fitting the original CIFAR-10 data via a Gaussian Mixture model. The new dataset will only keep the data points which have high probabilities to follow the Gaussian mixture distribution. We conduct two sets of adversarial training experiments using ResNet-18 model [He et al., 2016b]: one picking only 2 classes (airplane vs automobile) from CIFAR-10 and the other picking all 10 classes in CIFAR-10. The results are given in Figure 3.

From Figure 3, we can observe that for models trained on GMM filtered data, the overfitting issue is much less severe compared to the model trained on the original data. Specifically, for 2-class experiments, the overfitting is essentially benign for GMM filtered data. This partially backup our theoretical results of benign overfitting for adversarial classifiers trained on subGaussian mixture data, and when such data distribution assumption is violated, overfitting can become harmful. Furthermore, while Rice et al. [2020] only presents the negative result on empirical data distributions, we actually present a positive result that benign overfitting can occur in adversarial training for certain data distributions. We believe that subGaussian mixtures would not be the only distribution that could lead to benign overfitting in robust classifiers, yet our study certainly advances the understanding toward overfitting in adversarial settings.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we show that the benign overfitting phenomenon can also occur in adversarial training. Specifically, we derive the risk bounds of the adversarially trained linear classifiers and show that under moderate ℓ_p-norm perturbations, they can achieve the near-optimal standard and adversarial risks, despite overfitting the noisy training data. The numerical experimental results also validate our theoretical findings. Our current analysis is limited to linear classifiers, while in practice, adversarial training is commonly used with neural networks. We believe our work is the first step towards analyzing benign overfitting in adversarially trained neural networks. Yet extending our current analysis to adversarially trained neural networks is highly non-trivial and we leave it as a future work.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. QG is supported in part by the National Science Foundation CAREER Award 1906169 and IIS-2008981, and the Sloan Research Fellowship. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

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

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2018.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019a.

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

Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *arXiv preprint arXiv:2104.13628*, 2021.

Yuan Cao, Zixiang Chen, Mikhail Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *arXiv preprint arXiv:2202.06526*, 2022.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *NeurIPS*, pages 11192–11203, 2019.

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

Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *arXiv preprint arXiv:2004.12019*, 2020.

Niladri S Chatterji and Philip M Long. Foolish crowds support benign overfitting. *Journal of Machine Learning Research*, 23(125):1–12, 2022.

Lin Chen, Yifei Min, Mingrui Zhang, and Amin Karbasi. More data can expand the generalization gap between adversarially robust and standard models. In *International Conference on Machine Learning*, pages 1670–1680. PMLR, 2020.

Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. Pac-learning in the presence of evasion adversaries. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

Chen Dan, Yuting Wei, and Pradeep Ravikumar. Sharp statistical guaratees for adversarially robust gaussian classification. In *International Conference on Machine Learning*, pages 2345–2355. PMLR, 2020.

Ilias Diakonikolas, Daniel M Kane, and Pasin Manurangsi. The complexity of adversarially robust proper learning of halfspaces with agnostic noise. *Advances in Neural Information Processing Systems*, 2020.

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

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.

Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1186–1195, 2018.

Spencer Frei, Niladri S Chatterji, and Peter Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pages 2668–2703. PMLR, 2022.

Ruiqi Gao, Tianle Cai, Haochuan Li, Cho-Jui Hsieh, Liwei Wang, and Jason D Lee. Convergence of adversarial training in overparametrized neural networks. *Advances in Neural Information Processing Systems*, 32:13029–13040, 2019.

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*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018a.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018b.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–645. Springer, 2016b.

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdelrahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29, 2012.

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

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019a.

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

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Neurips*, pages 1097–1105, 2012.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

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

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

Chen Liu, Zhichao Huang, Mathieu Salzmann, Tong Zhang, and Sabine Süsstrunk. On the impact of hard adversarial instances on overfitting in adversarial training. *arXiv preprint arXiv:2112.07324*, 2021.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ICML*, 2018.

Omar Montasser, Surbhi Goel, Ilias Diakonikolas, and Nathan Srebro. Efficiently learning adversarially robust halfspaces with noise. In *International Conference on Machine Learning*, pages 7010–7021. PMLR, 2020.

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*, pages 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 J Zico Kolter. Overfitting in adversarially robust deep learning. *ICML*, 2020.

Amartya Sanyal, Puneet K. Dokania, Varun Kanade, and Philip Torr. How benign is benign overfitting ? In *International Conference on Learning Representations*, 2021.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Mądry. Adversarially robust generalization requires more data. *Advances in Neural Information Processing Systems*, 2018.

Ohad Shamir. The implicit bias of benign overfitting. *arXiv preprint arXiv:2201.11489*, 2022.

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.

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

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

Yoav Wald, Gal Yona, Uri Shalit, and Yair Carmon. Malign overfitting: Interpolation can provably preclude invariance. *arXiv preprint arXiv:2211.15724*, 2022.

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)*, pages 4030–4034. IEEE, 2021.

Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. In *ICML*, pages 6586–6595, 2019.

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*, 2020.

Boxi Wu, Jinghui Chen, Deng Cai, Xiaofei He, and Quanquan Gu. Do wider neural networks really help adversarial robustness? *arXiv preprint arXiv:2010.01279*, 2020.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33, 2020.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. *Advances in Neural Information Processing Systems*, 33, 2020.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pages 7085–7094. PMLR, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, pages 7472–7482, 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*, pages 3883–3893. PMLR, 2020a.

Yi Zhang, Orestis Plevrakis, Simon S Du, Xingguo Li, Zhao Song, and Sanjeev Arora. Over-parameterized adversarial training: An analysis overcoming the curse of dimensionality. *Advances in Neural Information Processing Systems*, 2020b.

Difan Zou, Spencer Frei, and Quanquan Gu. Provable robustness of adversarial training for learning halfspaces with noise. *International Conference on Machine Learning*, 2021a.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Benign overfitting of constant-stepsize sgd for linear regression. In *Conference on Learning Theory*, pages 4633–4635. PMLR, 2021b.