Benign Overfitting in Adversarially Robust Linear Classification (Supplementary Material)

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

A COMPARISON WITH DAN ET AL. (2020), TAHERI ET AL. (2020) AND JAVANMARD & SOLTANOLKOTABI (2020)

Dan et al. [2020] proposed an adversarial signal to noise ratio and studied the excess risk lower/upper bounds for learning Gaussian mixture models. Compared to the setting studied in Dan et al. [2020], our setting covers additional label flipping noises. More importantly, we study an estimator found by gradient descent that overfits the training data, while Dan et al. [2020] studied a specific plug-in estimator which does not overfit the training data. Due to these differences, there is a discrepancy in the risk bounds derived in both papers.

Taheri et al. [2020], Javanmard and Soltanolkotabi [2020] studied adversarial learning of linear models in the proportional limit setting, i.e., $d/n = O(1)$. In this setting, the data Gram matrix and the sample covariance matrix can be studied based on random matrix theory/Gaussian comparison inequalities/convex Gaussian min-max theorem. In contrast, in our setting where $d > \tilde{O}(n^2)$, the sample covariance matrix is singular but the $n \times n$ Gram matrix concentrates around its expectation. Therefore, our setting is different from the proportional limit setting in Taheri et al. [2020], Javanmard and Soltanolkotabi [2020], and these results are not directly comparable.

B PROOF OF KEY TECHNICAL LEMMAS

B.1 PROOF OF LEMMA ??

Proof. We first prove that $L(\boldsymbol{\theta}_1) \le 2n$. To show this, we observe that $\boldsymbol{\theta}_1 = \alpha_0 \sum_{k=1}^n \mathbf{z}_k$. Therefore

$$
\begin{aligned}
L(\boldsymbol{\theta}_1) &= \sum_{k=1}^n \exp(-\boldsymbol{\theta}_1^\top \mathbf{z}_k + \epsilon \|\boldsymbol{\theta}_1\|_q) \\
&= \sum_{k=1}^n \exp\bigg(-\alpha_0 \sum_{i=1}^n \mathbf{z}_i^\top \mathbf{z}_k + \alpha_0 \epsilon \bigg\| \sum_{i=1}^n \mathbf{z}_i \bigg\|_q \bigg) \\
&\le \sum_{k=1}^n \exp\bigg(\alpha_0 n \Big(c_0 \big(\|\boldsymbol{\mu}\|_2^2 + \sqrt{d \log(n/\delta)} \big) + \epsilon \sqrt{c_0} d \Big) \bigg) \\
&\le \sum_{k=1}^n \exp(1/16) \le 2n,
\end{aligned}
$$

where the first equality holds due to Lemma **??** and the fact that for any $\mathbf{u} \in \mathbb{R}^d$, $\|\mathbf{u}\|_q \le \|\mathbf{u}\|_1 \le \sqrt{d}\|\mathbf{u}\|_2$, while the second inequality is by the choice of sufficiently small α_0 and the assumptions that $d \ge Cn\|\boldsymbol{\mu}\|_2^2$ and $\epsilon \le R$ for some absolute constants C and R.

The rest part of Lemma **??** summarizes parts of the results in Li et al. [2020]. However, the results in Li et al. [2020] are

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

derived under the setting that $\|\mathbf{x}_i\|_2 \le 1$, Therefore to prove lemma **??**, we re-scale our data and model parameters and convert our setting to the setting in Li et al. [2020].

By lemma **??**, with probability at least $1-\delta$, $\|\mathbf{x}_i\|_2^2 \le c_0 d$ for all $i \in [n]$. We therefore denote $B := \sqrt{c_0 d}$, and then $\tilde{\mathbf{x}}_i := \mathbf{x}_i / B$ has ℓ_2-norm less than or equal to one. Further denote by $\boldsymbol{\beta}_t$ the linear model parameters in Li et al. [2020]'s algorithm, $\tilde{\mathbf{z}}_i = y_i \tilde{\mathbf{x}}_i$, η_t as their step sizes, $\tilde{\epsilon}$ as their perturbation strength, and

$$\tilde{\gamma} := \max_{\|\boldsymbol{\theta}\|_2=1} \min_{i\in[n]} y_i \boldsymbol{\theta}^\top \tilde{\mathbf{x}}_i$$

as the ℓ_p margin. Then the adversarial training update rule in Li et al. [2020] is

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \frac{\eta_t}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\beta}} \exp(-\boldsymbol{\beta}_t^\top \tilde{\mathbf{z}}_k + \tilde{\epsilon}\|\boldsymbol{\beta}_t\|_q).$$

Note that our update rule is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha_t \sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_k + \epsilon\|\boldsymbol{\theta}_t\|_q).$$

Now, in order to apply the results in Li et al. [2020], we convert our parameters to match their scaling. Since

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \alpha_t \sum_i \nabla_{\boldsymbol{\theta}} \exp(-B\boldsymbol{\theta}_t^\top \mathbf{z}_k/B + \epsilon\|B\boldsymbol{\theta}_t\|_q/B) \\ &= \boldsymbol{\theta}_t - \frac{nB\alpha_t}{n} \sum_i \nabla_{(B\boldsymbol{\theta})} \exp(-B\boldsymbol{\theta}_t^\top \mathbf{z}_k/B + \epsilon\|B\boldsymbol{\theta}_t\|_q/B).\end{aligned}$$

Therefore

$$B\boldsymbol{\theta}_{t+1} = B\boldsymbol{\theta}_t - \frac{nB^2\alpha_t}{n} \sum_i \nabla_{(B\boldsymbol{\theta})} \exp(-B\boldsymbol{\theta}_t^\top \mathbf{z}_k/B + \epsilon\|B\boldsymbol{\theta}_t\|_q/B).$$

It is easy to observe that we can now apply Theorem 3.3 and Theorem 3.4 in Li et al. [2020] by setting $\boldsymbol{\beta}_t = B\boldsymbol{\theta}_t, \eta_t = nB^2\alpha_t, \tilde{\epsilon} = \epsilon/B$. Moreover, by $\tilde{\mathbf{x}}_i = \mathbf{x}_i/B$, $\tilde{\epsilon} = \epsilon/B$ and the definition of $\tilde{\gamma}$, we have $\tilde{\gamma} = \bar{\gamma}/B$. Based on these relations, it is easy to see that under the conditions of Lemma **??**, $\tilde{\mathbf{x}}_i$, η_t, $\tilde{\epsilon}$, $\tilde{\gamma}$ satisfy the assumptions of Theorems 3.3 and 3.4 in Li et al. [2020]. Now (**??**) is an intermediate result of the proof of Theorem 3.3 in Li et al. [2020], and (**??**) follows by Theorem 3.4 in Li et al. [2020]. ☐

B.2 PROOF OF LEMMA ??

Proof. We have

$$\begin{aligned}\|\boldsymbol{\theta}_{t+1}\|_2 &= \left\| \sum_{m=0}^{t} \alpha_m \cdot \nabla L(\boldsymbol{\theta}_m) \right\|_2 \\ &\le \sum_{m=0}^{t} \alpha_m \|\nabla L(\boldsymbol{\theta}_m)\|_2 \\ &\le \sum_{m=0}^{t} \alpha_m \left\| \sum_{k=1}^{n} \big(\mathbf{z}_k - \epsilon \cdot \partial\|\boldsymbol{\theta}_m\|_q\big) \cdot \exp\big(-\mathbf{z}_k^\top \boldsymbol{\theta}_m + \epsilon\|\boldsymbol{\theta}_m\|_q\big) \right\|_2,\end{aligned}$$

where the first three inequalities hold by triangle inequality. By Lemma 2, we have

$$\begin{aligned}\|\boldsymbol{\theta}_{t+1}\|_2 &\le \sum_{m=0}^{t} \alpha_m \sum_{k=1}^{n} (\|\mathbf{z}_k\|_2 + \epsilon\sqrt{d}) \cdot \exp\big(-\mathbf{z}_k^\top \boldsymbol{\theta}_m + \epsilon\|\boldsymbol{\theta}_m\|_q\big) \\ &\le (\sqrt{c_0} + \epsilon)\sqrt{d} \sum_{m=0}^{t} \alpha_m \sum_{k=1}^{n} \cdot \exp\big(-\mathbf{z}_k^\top \boldsymbol{\theta}_m + \epsilon\|\boldsymbol{\theta}_m\|_q\big) \\ &= (\sqrt{c_0} + \epsilon)\sqrt{d} \sum_{m=0}^{t} \alpha_m L(\boldsymbol{\theta}_m),\end{aligned}$$

where the second inequality is due to Lemma **??**. ☐

B.3 PROOF OF LEMMA ??

Proof. Denote $E_k^t = \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_k)$ and $A_t^{i,j} = E_i^t / E_j^t$. The goal is to show that $\max_{i,j} A_t^{i,j} \le c_3$ for some constant $c_3 = 5c_0^2$. We prove this by induction.

For the base case ($t = 0$), we have $E_k^0 = \exp(0) = 1$. Therefore we have $\max_{i,j} A_0^{i,j} = 1 \le 5c_0^2$.

For $t > 0$, to simplify the notation, let E_1^t and E_2^t denote values for the first and second samples and their ratio $A_t := E_1^t / E_2^t$. We want to show that $A_{t+1} \le 5c_0^2$ (note that a similar result can be obtained for any distinct pair thus the max also satisfies).

Notice that

$$
\begin{aligned}
A_{t+1} &= \frac{\exp(-\boldsymbol{\theta}_{t+1}^\top \mathbf{z}_1)}{\exp(-\boldsymbol{\theta}_{t+1}^\top \mathbf{z}_2)} = \frac{\exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_1)}{\exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_2)} \cdot \frac{\exp(\alpha_t \nabla L(\boldsymbol{\theta}_t)^\top \mathbf{z}_1)}{\exp(\alpha_t \nabla L(\boldsymbol{\theta}_t)^\top \mathbf{z}_2)} \\
&= A_t \cdot \frac{\exp(-\alpha_t \sum_{k=1}^n (\mathbf{z}_k - \epsilon \partial \|\boldsymbol{\theta}_t\|_q)^\top \mathbf{z}_1 \cdot \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_k + \epsilon \|\boldsymbol{\theta}_t\|_q))}{\exp(-\alpha_t \sum_{k=1}^n (\mathbf{z}_k - \epsilon \partial \|\boldsymbol{\theta}_t\|_q)^\top \mathbf{z}_2 \cdot \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_k + \epsilon \|\boldsymbol{\theta}_t\|_q))} \\
&= A_t \cdot \underbrace{\frac{\exp(-\alpha_t (\mathbf{z}_1 - \epsilon \partial \|\boldsymbol{\theta}_t\|_q)^\top \mathbf{z}_1 \cdot \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_k + \epsilon \|\boldsymbol{\theta}_t\|_q))}{\exp(-\alpha_t (\mathbf{z}_2 - \epsilon \partial \|\boldsymbol{\theta}_t\|_q)^\top \mathbf{z}_2 \cdot \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_k + \epsilon \|\boldsymbol{\theta}_t\|_q))}}_{I_1} \\
&\quad \cdot \underbrace{\frac{\exp(-\alpha_t \sum_{k \ne 1}^n (\mathbf{z}_k - \epsilon \partial \|\boldsymbol{\theta}_t\|_q)^\top \mathbf{z}_1 \cdot \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_k + \epsilon \|\boldsymbol{\theta}_t\|_q))}{\exp(-\alpha_t \sum_{k \ne 2}^n (\mathbf{z}_k - \epsilon \partial \|\boldsymbol{\theta}_t\|_q)^\top \mathbf{z}_2 \cdot \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_k + \epsilon \|\boldsymbol{\theta}_t\|_q))}}_{I_2}.
\end{aligned} \tag{1}
$$

For term I_1, note that by Lemma **??** we have

$$
\sqrt{\frac{d}{c_0}} \le \|\mathbf{z}_k\|_2 \le \sqrt{c_0} d.
$$

Also since by Lemma 2, we have $\big\|\partial \|\boldsymbol{\theta}_t\|_q\big\|_p = 1$,

$$
|\mathbf{z}_k^\top \partial \|\boldsymbol{\theta}_t\|_q| \le \|\mathbf{z}_k\|_q \cdot \big\|\partial \|\boldsymbol{\theta}_t\|_q\big\|_p = \|\mathbf{z}_k\|_q \le \|\mathbf{z}_k\|_1 \le \sqrt{d} \|\mathbf{z}_k\|_2 \le \sqrt{c_0} d. \tag{2}
$$

Therefore, we have

$$
\begin{aligned}
I_1 &\le \exp\bigg(-\alpha_t \Big(\frac{d}{c_0} - \epsilon \sqrt{c_0} d\Big) \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_1 + \epsilon \|\boldsymbol{\theta}_t\|_q) + \alpha_t \Big(c_0 d + \epsilon \sqrt{c_0} d\Big) \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_2 + \epsilon \|\boldsymbol{\theta}_t\|_q) \bigg) \\
&= \exp\bigg(-\alpha_t E_2^t \Big(\Big(\frac{d}{c_0} - \epsilon \sqrt{c_0} d\Big) A_t - \Big(c_0 d + \epsilon \sqrt{c_0} d\Big) \Big) \exp\big(\epsilon \|\boldsymbol{\theta}_t\|_q\big) \bigg).
\end{aligned} \tag{3}
$$

For term I_2, by (**??**) and (2) we have

$$
\begin{aligned}
I_2 &\le \exp\bigg(\alpha_t \Big(c_0 \big(\|\boldsymbol{\mu}\|_2^2 + \sqrt{d \log(n/\delta)}\big) + \epsilon \sqrt{c_0} d \Big) \Big(\sum_{k \ne 1}^n \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_k + \epsilon \|\boldsymbol{\theta}_t\|_q) + \sum_{k \ne 2}^n \exp(-\boldsymbol{\theta}_t^\top \mathbf{z}_k + \epsilon \|\boldsymbol{\theta}_t\|_q) \Big) \bigg) \\
&\le \exp\bigg(2\alpha_t L(\boldsymbol{\theta}_t) \Big(c_0 \big(\|\boldsymbol{\mu}\|_2^2 + \sqrt{d \log(n/\delta)}\big) + \epsilon \sqrt{c_0} d \Big) \bigg)
\end{aligned} \tag{4}
$$

Substitute (3) and (4) into (1), we have

$$
\begin{aligned}
A_{t+1} \le A_t \cdot \exp\bigg(-\alpha_t E_2^t \Big(\Big(\frac{d}{c_0} - \epsilon \sqrt{c_0} d\Big) A_t - \Big(c_0 d + \epsilon \sqrt{c_0} d\Big) \Big) \exp\big(\epsilon \|\boldsymbol{\theta}_t\|_q\big) \bigg) \\
\cdot \exp\bigg(2\alpha_t L(\boldsymbol{\theta}_t) \Big(c_0 \big(\|\boldsymbol{\mu}\|_2^2 + \sqrt{d \log(n/\delta)}\big) + \epsilon \sqrt{c_0} d \Big) \bigg).
\end{aligned} \tag{5}
$$

Let us consider two cases here. If $(d/c_0 - \epsilon\sqrt{c_0}d)A_t - (c_0 d + \epsilon\sqrt{c_0}d) > c_0 d$, i.e., $A_t > (2c_0 + \epsilon\sqrt{c_0})/(1/c_0 - \epsilon\sqrt{c_0})$, we further have

$$
\begin{aligned}
A_{t+1} &\leq A_t \cdot \exp\Big(- \alpha_t E_2^t c_0 d \exp\big(\epsilon\|\boldsymbol{\theta}_t\|_q\big)\Big) \cdot \exp\bigg(2\alpha_t L(\boldsymbol{\theta}_t)\Big(c_0\big(\|\boldsymbol{\mu}\|_2^2 + \sqrt{d\log(n/\delta)}\big) + \epsilon\sqrt{c_0}d\Big)\bigg) \\
&\leq A_t \cdot \exp\Big(- \alpha_t E_2^t c_0 d \exp\big(\epsilon\|\boldsymbol{\theta}_t\|_q\big)\Big) \\
&\qquad \cdot \exp\bigg(2\alpha_t n E_2^t\Big(c_0\big(\|\boldsymbol{\mu}\|_2^2 + \sqrt{d\log(n/\delta)}\big) + \epsilon\sqrt{c_0}d\Big) \exp\big(\epsilon\|\boldsymbol{\theta}_t\|_q\big)\bigg) \\
&= A_t \cdot \exp\Big(- \alpha_t E_2^t c_0 \big(d - 2n\|\boldsymbol{\mu}\|_2^2 - 2n\sqrt{d\log(n/\delta)} - 2n\epsilon\sqrt{c_0}\big) \exp\big(\epsilon\|\boldsymbol{\theta}_t\|_q\big)\Big) \\
&\leq A_t,
\end{aligned}
$$

where the second inequality is due to the fact that $L(\boldsymbol{\theta}_t) = \sum_{k=1}^n E_k^t \exp\big(\epsilon\|\boldsymbol{\theta}_t\|_q\big)$ and $E_2^t = \max_k E_k^t$ while the last inequality holds since $d \geq C \cdot \max\{n\|\boldsymbol{\mu}\|_2^2, n^2\log(n/\delta)\}$.

On the other hand, if $A_t \leq (2c_0 + \epsilon\sqrt{c_0})/(1/c_0 - \epsilon\sqrt{c_0})$, we have

$$
\begin{aligned}
A_{t+1} &\leq A_t \cdot \exp\Big(\alpha_t E_2^t\big(c_0 d + \epsilon\sqrt{c_0}d\big) \exp\big(\epsilon\|\boldsymbol{\theta}_t\|_q\big)\Big) \\
&\qquad \cdot \exp\bigg(2\alpha_t L(\boldsymbol{\theta}_t)\Big(c_0\big(\|\boldsymbol{\mu}\|_2^2 + \sqrt{d\log(n/\delta)}\big) + \epsilon\sqrt{c_0}d\Big)\bigg) \\
&\leq A_t \cdot \exp\Big(\alpha_t L(\boldsymbol{\theta}_t)\big(c_0 d + \epsilon\sqrt{c_0}d\big)\Big) \cdot \exp\bigg(2\alpha_t L(\boldsymbol{\theta}_t)\Big(c_0\big(\|\boldsymbol{\mu}\|_2^2 + \sqrt{d\log(n/\delta)}\big) + \epsilon\sqrt{c_0}d\Big)\bigg) \\
&\leq A_t \cdot \exp\bigg(2\alpha_t n\Big(c_0\big(2\|\boldsymbol{\mu}\|_2^2 + 2\sqrt{d\log(n/\delta)} + d\big) + 3\epsilon\sqrt{c_0}d\Big)\bigg) \\
&\leq (2c_0 + \epsilon\sqrt{c_0})/(1/c_0 - \epsilon\sqrt{c_0}) \cdot \exp(1/8) \\
&\leq 5c_0^2,
\end{aligned}
$$

where the first inequality is due to the fact that $A_t > 0$, the third inequality holds by Lemma **??**, the fourth inequality is because $\alpha_t \leq 1/(c_0 C n d)$ and $d \geq C \cdot \max\{n\|\boldsymbol{\mu}\|_2^2, n^2\log(n/\delta)\}$ and the last inequality is because $\epsilon < C'$ and C' can be chosen such that $C' \leq 1/(2c_0^{1.5})$ and we have $1/c_0 - \epsilon\sqrt{c_0} > 1/(2c_0)$.

This concludes the proof. $\square$

B.4 PROOF OF LEMMA ??

Proof. Note that

$$
\begin{aligned}
\boldsymbol{\mu}^\top\boldsymbol{\theta}_{t+1} &= \boldsymbol{\mu}^\top\Big(\boldsymbol{\theta}_t + \alpha_t\sum_{k=1}^n \big(\mathbf{z}_k - \epsilon\partial\|\boldsymbol{\theta}_t\|_q\big)\exp(-\boldsymbol{\theta}_t^\top\mathbf{z}_k + \epsilon\|\boldsymbol{\theta}\|_1)\Big) \\
&= \boldsymbol{\mu}^\top\boldsymbol{\theta}_t - \alpha_t\epsilon\cdot\boldsymbol{\mu}^\top\partial\|\boldsymbol{\theta}_t\|_q\cdot L(\boldsymbol{\theta}_t) + \alpha_t\sum_{k=1}^n \big(\boldsymbol{\mu}^\top\mathbf{z}_k\big)\exp(-\boldsymbol{\theta}_t^\top\mathbf{z}_k + \epsilon\|\boldsymbol{\theta}\|_q)\Big) \\
&\geq \boldsymbol{\mu}^\top\boldsymbol{\theta}_t - \alpha_t\epsilon\|\boldsymbol{\mu}\|_q\cdot L(\boldsymbol{\theta}_t) + \alpha_t\sum_{k\in\mathcal{C}} \big(\boldsymbol{\mu}^\top\mathbf{z}_k\big)\exp(-\boldsymbol{\theta}_t^\top\mathbf{z}_k + \epsilon\|\boldsymbol{\theta}\|_q)\Big) \\
&\qquad + \alpha_t\sum_{k\in\mathcal{N}} \big(\boldsymbol{\mu}^\top\mathbf{z}_k\big)\exp(-\boldsymbol{\theta}_t^\top\mathbf{z}_k + \epsilon\|\boldsymbol{\theta}\|_q)\Big),
\end{aligned}
\tag{6}
$$

where the inequality holds in the same way as in (2). By Lemma **??** ((**??**) and (**??**)), we further bound (6) by

$$\begin{aligned}\boldsymbol{\mu}^\top\boldsymbol{\theta}_{t+1} &\geq \boldsymbol{\mu}^\top\boldsymbol{\theta}_t - \alpha_t\epsilon\|\boldsymbol{\mu}\|_q \cdot L(\boldsymbol{\theta}_t) + \frac{\alpha_t}{2}\sum_{k\in\mathcal{C}}\|\boldsymbol{\mu}\|_2^2\exp(-\boldsymbol{\theta}_t^\top\mathbf{z}_k + \epsilon\|\boldsymbol{\theta}\|_q)\Big) \\ &\quad - \frac{3\alpha_t}{2}\sum_{k\in\mathcal{N}}\|\boldsymbol{\mu}\|_2^2\exp(-\boldsymbol{\theta}_t^\top\mathbf{z}_k + \epsilon\|\boldsymbol{\theta}\|_q)\Big) \\ &= \boldsymbol{\mu}^\top\boldsymbol{\theta}_t - \alpha_t\epsilon\|\boldsymbol{\mu}\|_q \cdot L(\boldsymbol{\theta}_t) + \frac{\alpha_t}{2}\|\boldsymbol{\mu}\|_2^2 L(\boldsymbol{\theta}_t) - 2\alpha_t\|\boldsymbol{\mu}\|_2^2\sum_{k\in\mathcal{N}}\exp(-\boldsymbol{\theta}_t^\top\mathbf{z}_k + \epsilon\|\boldsymbol{\theta}\|_q)\Big). \end{aligned} \tag{7}$$

Note that we have

$$\begin{aligned}\sum_{k\in\mathcal{N}}\exp(-\boldsymbol{\theta}_t^\top\mathbf{z}_k + \epsilon\|\boldsymbol{\theta}\|_q) &= \sum_{k\in\mathcal{N}}\exp(-\boldsymbol{\theta}_t^\top\mathbf{z}_k)\cdot\exp(\epsilon\|\boldsymbol{\theta}\|_q) \\ &\leq c_3(\eta + c_1)n \cdot \Big(\max_k E_k\Big)\cdot\exp(\epsilon\|\boldsymbol{\theta}\|_q) \\ &\leq c_3(\eta + c_1)L(\boldsymbol{\theta}_t) \\ &\leq \frac{1}{8}L(\boldsymbol{\theta}_t),\end{aligned}$$

where the first inequality is due to Lemma **??** and the last inequality is because $\eta < 1/C$ and c_1 can be chosen arbitrarily small given sufficient large C. Therefore, (7) can be further written as

$$\begin{aligned}\boldsymbol{\mu}^\top\boldsymbol{\theta}_{t+1} &\geq \boldsymbol{\mu}^\top\boldsymbol{\theta}_t - \alpha_t\epsilon\|\boldsymbol{\mu}\|_q \cdot L(\boldsymbol{\theta}_t) + \frac{\alpha_t}{2}\|\boldsymbol{\mu}\|_2^2 L(\boldsymbol{\theta}_t) - \frac{\alpha_t}{4}\|\boldsymbol{\mu}\|_2^2 L(\boldsymbol{\theta}_t) \\ &= \boldsymbol{\mu}^\top\boldsymbol{\theta}_t + \alpha_t\bigg(\frac{\|\boldsymbol{\mu}\|_2^2}{4} - \epsilon\|\boldsymbol{\mu}\|_q\bigg)\cdot L(\boldsymbol{\theta}_t) \\ &= \bigg(\frac{\|\boldsymbol{\mu}\|_2^2}{4} - \epsilon\|\boldsymbol{\mu}\|_q\bigg)\cdot\sum_{m=0}^{t}\alpha_m L(\boldsymbol{\theta}_m),\end{aligned} \tag{8}$$

where the last equality is due the fact that $\boldsymbol{\theta}_0 = \mathbf{0}$. Now we multiply $\|\mathbf{w}\|_2/\|\boldsymbol{\theta}_{t+1}\|_2$ on both sides of (8) and take $t \to \infty$

$$\lim_{t\to\infty}\frac{\|\mathbf{w}\|_2(\boldsymbol{\mu}^\top\boldsymbol{\theta}_{t+1})}{\|\boldsymbol{\theta}_{t+1}\|_2} \geq \lim_{t\to\infty}\bigg(\frac{\|\boldsymbol{\mu}\|_2^2}{4} - \epsilon\|\boldsymbol{\mu}\|_q\bigg)\frac{\|\mathbf{w}\|_2}{\|\boldsymbol{\theta}_{t+1}\|_2}\cdot\sum_{m=0}^{t}\alpha_m L(\boldsymbol{\theta}_m).$$

Since $\|\mathbf{w}\|_2 = 1$, and by Lemma **??**, it is easy to observe that $\mathbf{w} = \lim_{t\to\infty}\boldsymbol{\theta}_t/\|\boldsymbol{\theta}_t\|_2$, we have

$$\begin{aligned}\boldsymbol{\mu}^\top\mathbf{w} &\geq \bigg(\frac{\|\boldsymbol{\mu}\|_2^2}{4} - \epsilon\|\boldsymbol{\mu}\|_q\bigg)\cdot\lim_{t\to\infty}\frac{\sum_{m=0}^{t}\alpha_m L(\boldsymbol{\theta}_m)}{\|\boldsymbol{\theta}_{t+1}\|_2} \\ &\geq \bigg(\frac{\|\boldsymbol{\mu}\|_2^2}{4} - \epsilon\|\boldsymbol{\mu}\|_q\bigg)\frac{1}{(\sqrt{c_0}+\epsilon)\sqrt{d}}.\end{aligned}$$

where the last inequality is due to Lemma **??**. Note that Lemma **??** also suggests that $\|\boldsymbol{\theta}_t/\|\boldsymbol{\theta}_t\|_2 - \mathbf{w}\|_2 \leq c_3\log n/\log t$, we have

$$\begin{aligned}\boldsymbol{\mu}^\top\mathbf{w} &= \boldsymbol{\mu}^\top\bigg(\mathbf{w} - \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|_2} + \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|_2}\bigg) \\ &\leq \|\boldsymbol{\mu}\|_2\cdot\bigg\|\mathbf{w} - \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|_2}\bigg\|_2 + \frac{\boldsymbol{\mu}^\top\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|_2} \\ &\leq \frac{c_3\|\boldsymbol{\mu}\|_2\log n}{\log t} + \frac{\boldsymbol{\mu}^\top\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|_2}.\end{aligned}$$

Therefore,

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

$\square$

C AUXILIARY LEMMAS

Theorem 1 (Proposition 5.10 in Vershynin [2010]). *Let $X_1, X_2, \ldots, X_n$ be independent centered sub-Gaussian random variables, and let $K = \max_i \|X_i\|_{\psi_2}$. Then for every $a = (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$ and for every $t > 0$, we have*

$$\mathbb{P}\Big(\Big| \sum_{i=1}^{n} a_i X_i \Big| > t \Big) \leq \exp\Big(- \frac{Ct^2}{K^2 \|a\|_2^2} \Big),$$

where $C > 0$ is a constant.

Lemma 2. *For any $\boldsymbol{\theta} \in \mathbb{R}^d$,*

$$\big\| \partial \|\boldsymbol{\theta}\|_q \big\|_2 \leq \sqrt{d}, \ \big\| \partial \|\boldsymbol{\theta}\|_q \big\|_p = 1.$$

Proof. Note that we have

$$(\partial \|\boldsymbol{\theta}\|_q)_i = \frac{\theta_i^{q-1}}{\|\boldsymbol{\theta}\|_q^{q-1}} \cdot \operatorname{sign}(\boldsymbol{\theta}),$$

and since for any vector $\mathbf{u} \in \mathbb{R}^d$, $\|\mathbf{u}\|_q \geq \|\mathbf{u}\|_\infty$, $\|\mathbf{u}\|_2 \leq \sqrt{d}\|\mathbf{u}\|_\infty$, we have

$$\big\| \partial \|\boldsymbol{\theta}\|_q \big\|_2 = \frac{\big\|\boldsymbol{\theta}^{\circ(q-1)}\big\|_2}{\|\boldsymbol{\theta}\|_q^{q-1}} \leq \frac{\sqrt{d}\|\boldsymbol{\theta}\|_\infty^{q-1}}{\|\boldsymbol{\theta}\|_q^{q-1}} \leq \sqrt{d},$$

where $\circ$ denotes element-wise power. This concludes the first part of the lemma. For the second part, by p-norm definition, we have

$$\big\| \partial \|\boldsymbol{\theta}\|_q \big\|_p = \frac{\big\|\boldsymbol{\theta}^{\circ(q-1)}\big\|_p}{\|\boldsymbol{\theta}\|_q^{q-1}} = \frac{1}{\|\boldsymbol{\theta}\|_q^{q-1}} \Big(\sum_{i=1}^{d} (\theta_i^{q-1})^p \Big)^{1/p} = \frac{1}{\|\boldsymbol{\theta}\|_q^{q-1}} \bigg(\Big(\sum_{i=1}^{d} \theta_i^q \Big)^{1/q} \bigg)^{q-1} = 1.$$

□

D ADDITIONAL EXPERIMENTS

In this section, we present the additional experiments covering more settings as well as more complex models such as 2-layer neural network.

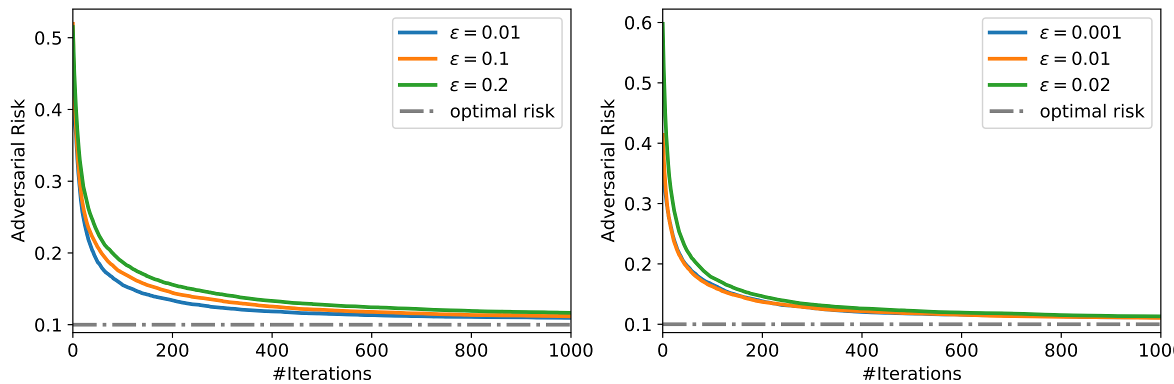
D.1 ADVERSARIALLY TRAINED LINEAR CLASSIFIER UNDER VARIOUS SETTINGS

In Figures 1,2,3, we plot the adversarial risk of adversarially trained linear classifiers versus the training iterations t for different perturbation level ϵ for various combinations of dimension d and $\|\boldsymbol{\mu}\|_2$. Specifically, in Figure 3, we can observe that with moderate perturbations and sufficient over-parameterization, adversarially trained linear classifiers can achieve near-optimal adversarial risk.

D.2 ADVERSARIALLY TRAINED 2-LAYER NEURAL NETWORKS

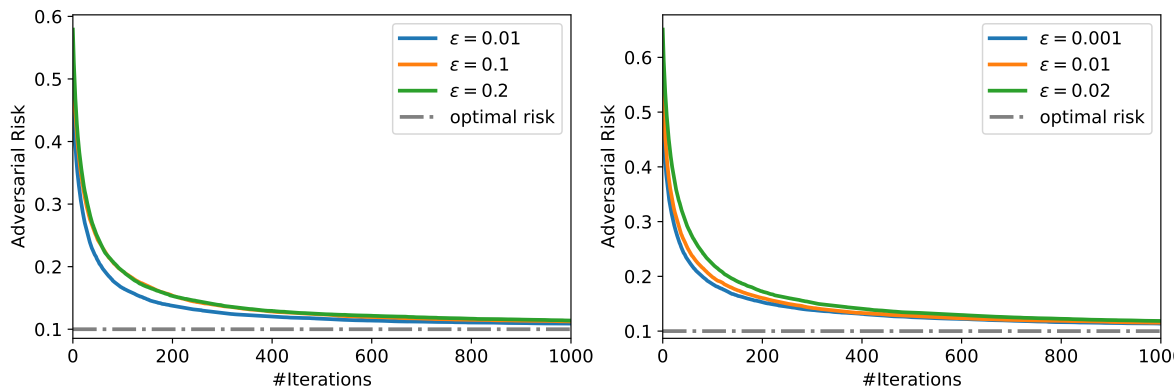
We have also conducted extra experiments on 2-layer neural networks with ReLU activation functions (one extra fix-dimension hidden layer). The data generation process are the same as our linear experiments. Note that in this setting, we no longer have the closed-form solutions to the inner maximization problem. Therefore, we following Madry et al. [2018] and use 10-step Projected Gradient Descent to get the inner maximizer.

As can be seen from Figure 4, the empirical results on 2-layer ReLU network suggest very similar trends as the linear classifier for both adversarial risk and standard risk. This further backs up our theoretical conclusions.



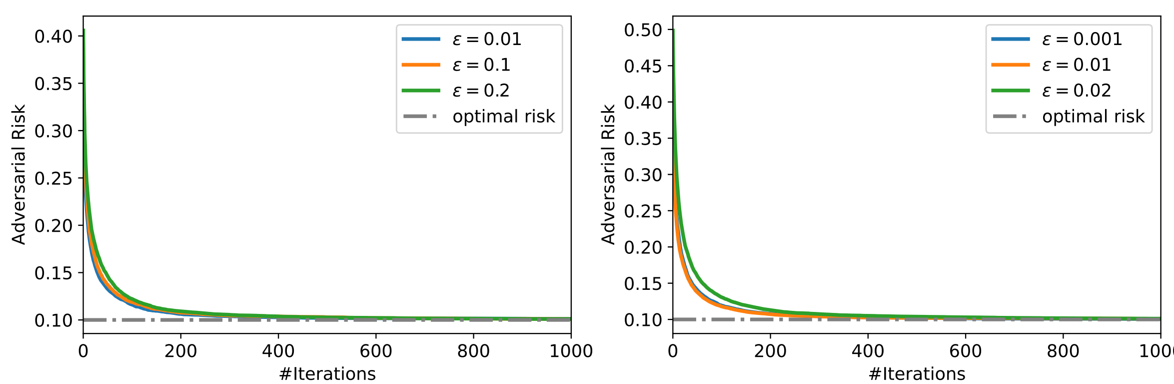
(a) ℓ_2 perturbation (b) ℓ_∞ perturbation

Figure 1: Risk and adversarial risk of adversarially trained linear classifiers versus the training iterations t for different perturbation level ϵ. The label noise level is set as $\eta = 0.1$, the training set size $n = 50$, dimension $d = 200$ and $\|\boldsymbol{\mu}\|_2 = d^{0.4}$. The train error reaches 0 for all experiments.



(a) ℓ_2 perturbation (b) ℓ_∞ perturbation

Figure 2: Risk and adversarial risk of adversarially trained linear classifiers versus the training iterations t for different perturbation level ϵ. The label noise level is set as $\eta = 0.1$, the training set size $n = 50$, dimension $d = 1000$ and $\|\boldsymbol{\mu}\|_2 = d^{0.3}$. The train error reaches 0 for all experiments.

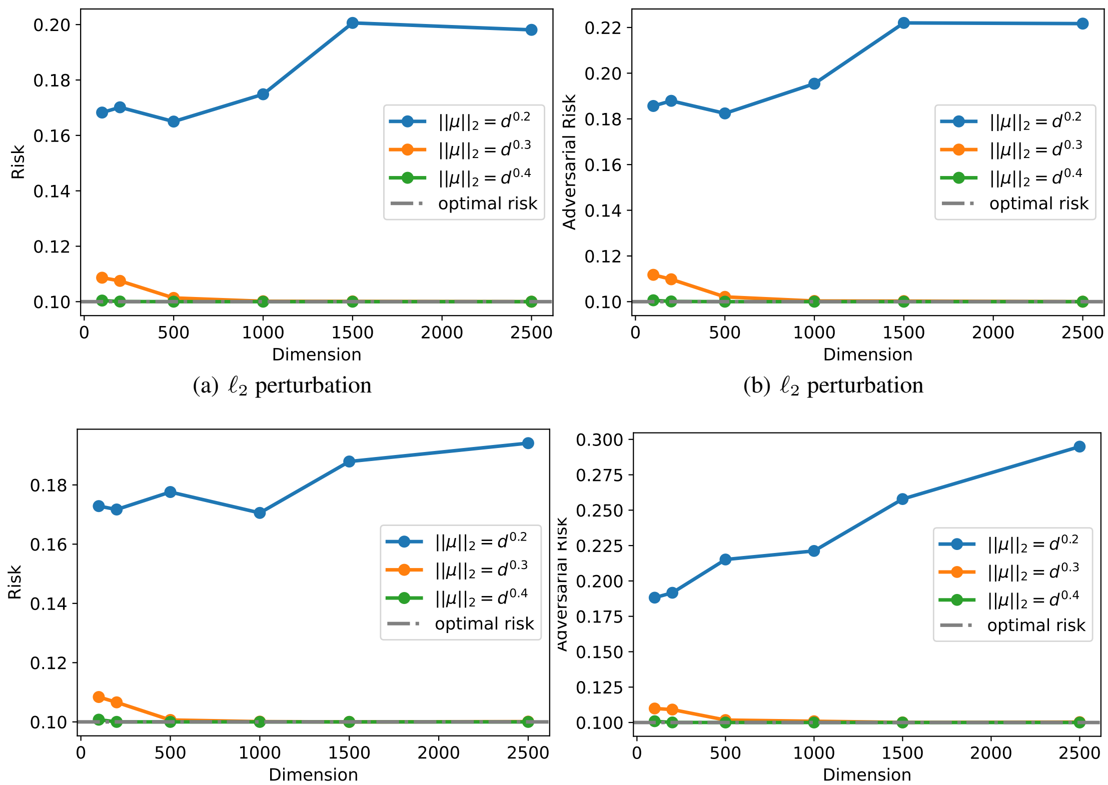


(a) ℓ_2 perturbation (b) ℓ_∞ perturbation

Figure 3: Risk and adversarial risk of adversarially trained linear classifiers versus the training iterations t for different perturbation level ϵ. The label noise level is set as $\eta = 0.1$, the training set size $n = 50$, dimension $d = 1000$ and $\|\boldsymbol{\mu}\|_2 = d^{0.4}$. The train error reaches 0 for all experiments.

References

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.



(a) ℓ_2 perturbation

(b) ℓ_2 perturbation

(c) ℓ_∞ perturbation

(d) ℓ_∞ perturbation

Figure 4: Risk and adversarial risk of adversarially trained 2-layer ReLU network 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.

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

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.

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

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

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.