The Curious Case of Adversarially Robust Models: More Data Can Help, Double Descend, or Hurt Generalization (Supplementary material)

Yifei Min[1] **Lin Chen**[2] **Amin Karbasi**[3]

[1]Dept. of Statistics and Data Science, Yale University, New Haven, CT, USA
[2]Simons Institute for the Theory of Computing, University of California, Berkeley, Berkeley, CA, USA
[3] Dept. of Electrical Engineering, Yale University, New Haven, CT, USA

1 PROOF OF RESULTS FOR THE GAUSSIAN MODEL

In this section we give the proof of Theorem 1 and Corollary 2.

Before proving Theorem 1, we need to establish several lemmas. First we restate the result by Chen et al. [2020b] that gives the closed form solution for the robust classifier.

Proposition 1 (Lemma 10 in Chen et al. [2020b]). *Given n training data points $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{\pm 1\}$ and $\varepsilon > 0$, if the robust classifier is defined as* (3)*, then we have $w_n^{\text{rob}} = W \operatorname{sign}(u - \varepsilon \operatorname{sign}(u))$, where $u = \frac{1}{n}\sum_{i=1}^n y_i x_i$.*

First, we define the error function $\text{erf}(\cdot) : \mathbb{R} \to \mathbb{R}$ by

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt\,, \tag{1}$$

and it has the following property.

Lemma 2. *If $z \sim \mathcal{N}(0,1)$, we have*

$$\mathbb{P}(z < x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right].$$

Proof of Lemma 2. In light of the density of the standard normal distribution and by a change of variable, we have

$$\mathbb{P}(z < x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}}\, dt = \frac{1}{2} + \frac{\sqrt{2}}{\sqrt{2\pi}} \int_0^{x/\sqrt{2}} e^{-s^2}\, ds = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right].$$

□

In addition, we define the function $L(\cdot,\cdot) :\ \mathbb{R}^2 \to \mathbb{R}$ by

$$L(v, \varepsilon') = \text{erf}\,(v) + \text{erf}\,(v\,(\varepsilon' - 1)) - \text{erf}\,(v\,(\varepsilon' + 1))\,. \tag{2}$$

For all $j \in [d]$, we define

$$v_j = \frac{\sqrt{n}\mu(j)}{\sqrt{2}\sigma(j)}\,, \quad \varepsilon'_j = \frac{\varepsilon}{\mu(j)}\,, \tag{3}$$

where $\mu(j)$ and $\sigma(j)$ are defined in the data generation process described at the beginning of Section 4.

Lemma 3 gives the expression for the generalization error.

Lemma 3. *Suppose that the generalization error is defined as in* (4)*. Then we have*

$$L_n = W \sum_{j\in[d]} \mu(j) L\left(v_j, \varepsilon'_j\right) \,,$$

where v_j and ε'_j are defined in (3)*.*

Proof of Lemma 3. By (4), Proposition 1 and the independence between test and training data, we have

$$\begin{aligned} L_n =& -\mathbb{E}_{\{(x_i,y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}_\mathcal{N}} \left[\mathbb{E}_{(x,y)\sim\mathcal{D}_\mathcal{N}}\left[y\langle w_n^{\text{rob}}, x\rangle\right]\right] = -\mathbb{E}_{\{(x_i,y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}_\mathcal{N}}\left[\langle w_n^{\text{rob}}, \mu\rangle\right] \\ =& -W\cdot \sum_{j\in[d]} \mu(j)\mathbb{E}_{\{(x_i,y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}_\mathcal{N}} \left[\text{sign}\left(u(j) - \varepsilon\,\text{sign}\left(u(j)\right)\right)\right] \end{aligned}$$

Since $y_i x_i \sim \mathcal{N}(\mu, \Sigma)$, we have $u \sim \mathcal{N}(\mu, \frac{\Sigma}{n})$, and it follows that

$$L_n = -W\cdot \sum_{j\in[d]} \mu(j)\mathbb{E}_{u(j)\sim\mathcal{N}(\mu(j), \frac{\sigma^2(j)}{n})} \left[\text{sign}\left(u(j) - \varepsilon\,\text{sign}\left(u(j)\right)\right)\right] \,.$$

Denote $I_j = -\mathbb{E}_{u(j)\sim\mathcal{N}\left(\mu(j), \frac{\sigma^2(j)}{n}\right)} \left[\text{sign}\left(u(j) - \varepsilon\,\text{sign}\left(u(j)\right)\right)\right]$. Then we have

$$\begin{aligned} I_j &= \mathbb{P}\left(u(j) < -\varepsilon\right) - \mathbb{P}\left(-\varepsilon < u(j) < 0\right) + \mathbb{P}\left(0 < u(j) < \varepsilon\right) - \mathbb{P}\left(\varepsilon < u(j)\right) \\ &= 1 - 2\mathbb{P}\left(-\varepsilon < u(j) < 0\right) - 2\mathbb{P}\left(\varepsilon < u(j)\right) \\ &= 1 - 2\mathbb{P}\left(\frac{(-\varepsilon-\mu(j))\sqrt{n}}{\sigma(j)} < z < \frac{-\mu(j)\sqrt{n}}{\sigma(j)}\right) - 2\mathbb{P}\left(\frac{(\varepsilon-\mu(j))\sqrt{n}}{\sigma(j)} < z\right) \\ &= 1 - 2\left[\mathbb{P}\left(z < \frac{(\varepsilon+\mu(j))\sqrt{n}}{\sigma(j)}\right) - \mathbb{P}\left(z < \frac{\mu(j)\sqrt{n}}{\sigma(j)}\right)\right] - 2\left[1 - \mathbb{P}\left(z < \frac{(\varepsilon-\mu(j))\sqrt{n}}{\sigma(j)}\right)\right] \,, \end{aligned}$$

where z is a standard normal random variable. By Lemma 2 we have

$$\begin{aligned} I_j =& \ \text{erf}\left(\frac{\mu(j)\sqrt{n}}{\sqrt{2}\sigma(j)}\right) + \text{erf}\left(\frac{(\varepsilon-\mu(j))\sqrt{n}}{\sqrt{2}\sigma(j)}\right) - \text{erf}\left(\frac{(\varepsilon+\mu(j))\sqrt{n}}{\sqrt{2}\sigma(j)}\right) \\ =& \ \text{erf}(v_j) + \text{erf}(v_j(\varepsilon'_j - 1)) - \text{erf}(v_j(\varepsilon'_j + 1)) = L(v_j, \varepsilon'_j) \,, \end{aligned}$$

which implies that $L_n = W\sum_{j\in[d]} \mu(j) L(v_j, \varepsilon'_j)$.

□

Note that $L(v, \varepsilon')$ is differentiable in v, and by our definition each v_j is smooth and monotonic in n. Together with Lemma 3 we know that L_n is differentiable w.r.t. n. Therefore, to study the dynamic of L_n in n, it is equivalent to studying the derivative $\frac{dL_n}{dn}$. We define the function $f(\cdot,\cdot) : \mathbb{R}^2 \to \mathbb{R}$ by

$$f\left(t, \varepsilon'\right) = t - (1+\varepsilon')t^{(1+\varepsilon')^2} - (1-\varepsilon')t^{(1-\varepsilon')^2} \,.$$

In Lemma 4, we compute the partial derivative of L.

Lemma 4. *Let $t = e^{-v^2}$ and f be defined as in* (1)*. The partial derivative of $L(v, \varepsilon')$ w.r.t. v is given by*

$$\frac{\partial L(v,\varepsilon')}{\partial v} = \frac{2}{\sqrt{\pi}} f(t, \varepsilon') \,.$$

Proof of Lemma 4. By (1) we have

$$\frac{d}{dx}\,\text{erf}(x) = \frac{2}{\sqrt{\pi}} e^{-x^2} \,,$$

and it follows by (2) that

$$\frac{\partial L(v,\varepsilon')}{\partial v} = \frac{2}{\sqrt{\pi}} e^{-v^2} + (\varepsilon'-1)\frac{2}{\sqrt{\pi}} e^{-v^2\cdot(\varepsilon'-1)^2} - (\varepsilon'+1)\frac{2}{\sqrt{\pi}} e^{-v^2\cdot(\varepsilon'+1)^2} = \frac{2}{\sqrt{\pi}} f(t, \varepsilon') \,.$$

□

The proof of Theorem 1 follows from studying the derivative $\frac{dL_n}{dn}$. Lemma 4 implies that the derivative depends on the sign of the function f. We investigate the sign of f in Lemma 5.

Lemma 5. *There exist $0 < \delta_1 \le \delta_2 < 1$ such that the following statements hold.*

(a) When $0 < \varepsilon' < \delta_1$, $f(t, \varepsilon') < 0$ for $\forall\, t \in (0, 1)$.
(b) When $\delta_2 < \varepsilon' < 1$, there exist $0 < \tau_1 < \tau_2 < 1$ depending on ε' such that

$$f(t, \varepsilon') \begin{cases} < 0 & \forall t \in (0, \tau_1)\,, \\ > 0 & \forall t \in (\tau_1, \tau_2)\,, \\ < 0 & \forall t \in (\tau_2, 1)\,, \end{cases}$$

and

$$\lim_{\varepsilon' \to 1^-} \tau_1(\varepsilon') = 0\,,$$
$$\tau_2(\varepsilon') \ge \frac{1}{3}\,.$$

(c) When $1 \le \varepsilon'$, $f(t, \varepsilon')$, there exists $\tau_2 < 1$ such that

$$f(t, \varepsilon') \begin{cases} > 0 & \forall t \in (0, \tau_2)\,, \\ < 0 & \forall t \in (\tau_2, 1)\,. \end{cases}$$

We compute the partial derivative of f w.r.t. t

$$f'(t, \varepsilon') = \frac{\partial f(t, \varepsilon')}{\partial t} = 1 - (1+\varepsilon')^3\, t^{(1+\varepsilon')^2 - 1} - (1-\varepsilon')^3\, t^{(1-\varepsilon')^2 - 1}\,.$$

The proof of Lemma 5 uses the following Lemma 6 and Lemma 7. To make it concise, whenever we fix ε' in the context, we omit ε' and write $f(t) = f(t, \varepsilon')$ and $f'(t) = f'(t, \varepsilon')$.

Lemma 6. *The right-sided limit of f' at 0 is given by*

$$\lim_{t \to 0^+} f'(t) = \begin{cases} -\infty & \text{if } 0 < \varepsilon' < 1\,, \\ 1 & \text{if } \varepsilon' = 1\,, \\ +\infty & \text{if } 1 < \varepsilon' < 2\,. \end{cases}$$

In addition, we have

$$\lim_{t \to 1^-} f'(t) < 0\,, \qquad \forall\, 0 < \varepsilon'\,.$$

The proof of Lemma 6 follows from direct computation. Using Lemma 6, we obtain Lemma 7.

Lemma 7. *For any fixed $0 < \varepsilon' < 1$, there exists some $t_0 = t_0(\varepsilon') \in (0, 1)$ such that $f'(t)$ is strictly increasing for $t \in (0, t_0)$ and strictly decreasing for $t \in (t_0, 1)$. For any fixed $1 \le \varepsilon' \le 2$, $f'(t)$ is strictly decreasing for $t \in (0, 1)$.*

Proof of Lemma 7. We differentiate f' w.r.t. t to get

$$\frac{\partial f'(t)}{\partial t} = -\,(1+\varepsilon')^3 \left[(1+\varepsilon')^2 - 1\right] t^{(1+\varepsilon')^2 - 2} - (1-\varepsilon')^3 \left[(1-\varepsilon')^2 - 1\right] t^{(1-\varepsilon')^2 - 2}\,.$$

First we consider the case where $0 < \varepsilon' < 1$. The function f' is continuously differentiable on $(t, \varepsilon') \in (0, 1) \times (0, 1)$. For any fixed $\varepsilon' < 1$, setting $\frac{\partial f'(t)}{\partial t} = 0$ yields the unique solution of t in $(0, 1)$ as

$$t_0 = \left[\left(\frac{1+\varepsilon'}{1-\varepsilon'}\right)^3 \left(\frac{2+\varepsilon'}{2-\varepsilon'}\right)\right]^{-\frac{1}{4\varepsilon'}}\,. \tag{4}$$

Since $\lim_{t\to 0^+} f'(t) = -\infty$, $f'(t)$ is strictly increasing w.r.t. $t \in (0, t_0)$. Also note that

$$\lim_{t\to 1^-} \frac{\partial f'(t)}{\partial t} = \lim_{t\to 1^-} -(1+\varepsilon')^3 \left[(1+\varepsilon')^2 - 1\right] t^{(1+\varepsilon')^2-2} - (1-\varepsilon')^3 \left[(1-\varepsilon')^2 - 1\right] t^{(1-\varepsilon')^2-2}$$
$$= -2\varepsilon'^2 \left(5\varepsilon'^2 + 7\right) < 0\,,$$

which together with $\frac{\partial}{\partial t}(f'(t_0)) = 0$ indicates that $f'(t)$ is strictly decreasing for $t \in (t_0, 1)$. We conclude that t_0 is the unique local extreme and also the global maximum of $f'(t)$ on $t \in (0,1)$.

For $1 \le \varepsilon' \le 2$, we have for all $t \in (0,1)$

$$-(1+\varepsilon')^3 \left[(1+\varepsilon')^2 - 1\right] t^{(1+\varepsilon')^2-2} < 0\,,$$
$$-(1-\varepsilon')^3 \left[(1-\varepsilon')^2 - 1\right] t^{(1-\varepsilon')^2-2} \le 0\,.$$

It follows that $\frac{\partial f'(t)}{\partial t} < 0$, which implies that $f'(t)$ is strictly decreasing.

□

A direct application of Lemma 7 gives the following Lemma 8

Lemma 8. *For all $0 < \varepsilon' < 1$ sufficiently close to 1, $f'(t)$ has exactly two zeros on $t \in (0,1)$.*

Proof of Lemma 8. By Lemma 7, we know that $f'(t)$ is strictly increasing on $t \in (0, t_0)$ and strictly decreasing on $(t_0, 1)$. Recall that Lemma 6 shows that for $0 < \varepsilon' < 1$, $\lim_{t\to 0^+} f'(t) = -\infty$ and $\lim_{t\to 1^-} f'(t) < 0$. Therefore it suffices to show $f'(t_0) > 0$ for all ε' sufficiently close to 1^-. We define

$$A = \left(\frac{1+\varepsilon'}{1-\varepsilon'}\right)^3 \left(\frac{2+\varepsilon'}{2-\varepsilon'}\right). \tag{5}$$

We have A tends to $+\infty$ as $\varepsilon' \to 1^-$. We then write

$$f'(t_0) = 1 - (1+\varepsilon')^3 A^{-\frac{1}{2}-\frac{\varepsilon'}{4}} - (1-\varepsilon')^3 A^{\frac{1}{2}-\frac{\varepsilon'}{4}}\,.$$

Note that $\lim_{\varepsilon'\to 1^-} (1+\varepsilon')^3 A^{-\frac{1}{2}-\frac{\varepsilon'}{4}} = 0$, and

$$\lim_{\varepsilon'\to 1^-} (1-\varepsilon')^3 A^{\frac{1}{2}-\frac{\varepsilon'}{4}} = \lim_{\varepsilon'\to 1^-} (1-\varepsilon')^{\frac{3}{2}+\frac{3\varepsilon'}{4}} \cdot \left[(1+\varepsilon')^3 \left(1 + \frac{2\varepsilon'}{2-\varepsilon'}\right)\right]^{\frac{1}{2}-\frac{\varepsilon'}{4}} = 0\,.$$

Therefore we conclude that $f'(t_0) > 0$ as $\varepsilon' \to 1^-$.

□

We denote the two zeros in Lemma 8 by $t_1 = t_1(\varepsilon')$ and $t_2 = t_2(\varepsilon')$ where $t_1 < t_2$.

Now we are ready to prove Lemma 5.

Proof of Lemma 5. We show (a) first. Note that for any fixed $\varepsilon' < 1$, $f(0) = 0$. Therefore it suffices to show that for any ε' sufficiently close to 0, the derivative $f'(t) < 0$. Since by Lemma 7 we have $f'(t) < \sup_{t\in(0,1)} f'(t) = f'(t_0)$ when $0 < \varepsilon' < 1$, it remains to show that $f'(t_0) < 0$ for all ε' sufficiently close to 0.

In light of (4), $f'(t_0) < 0$ is equivalent to

$$1 - (1+\varepsilon')^3 \left[\left(\frac{1+\varepsilon'}{1-\varepsilon'}\right)^3 \left(\frac{2+\varepsilon'}{2-\varepsilon'}\right)\right]^{-\frac{\varepsilon'^2+2\varepsilon'}{4\varepsilon'}} - (1-\varepsilon')^3 \left[\left(\frac{1+\varepsilon'}{1-\varepsilon'}\right)^3 \left(\frac{2+\varepsilon'}{2-\varepsilon'}\right)\right]^{-\frac{\varepsilon'^2-2\varepsilon'}{4\varepsilon'}} < 0\,.$$

Recall that we define

$$A = \left(\frac{1+\varepsilon'}{1-\varepsilon'}\right)^3 \left(\frac{2+\varepsilon'}{2-\varepsilon'}\right).$$

Rearranging the terms yields $A^{\varepsilon'/4} < (1+\varepsilon')^3 A^{-1/2} + (1-\varepsilon')^3 A^{1/2}$. Since $A > 1$ and $\varepsilon' < 1$, we have $A^{\varepsilon'/4} < A^{1/2}$. Thus it now suffices to show $A^{1/2} < (1+\varepsilon')^3 A^{-1/2} + (1-\varepsilon')^3 A^{1/2}$, or equivalently $A < (1+\varepsilon')^3/[1-(1-\varepsilon')^3]$. We can further simplify this into

$$\frac{2+\varepsilon'}{2-\varepsilon'} < \frac{(1-\varepsilon')^3}{1-(1-\varepsilon')^3} .$$

Finally, note that LHS $\to 1$ and RHS $\to +\infty$ as $\varepsilon' \to 0^+$. Therefore there must exist $\delta_1 \in (0,1)$ such that: for any $0 < \varepsilon' < \delta_1$, $f'(t) < 0$ for all $t \in (0,1)$. Thus $f(t) < 0$ for all $t \in (0,1)$.

Now we show (b). By Lemma 8, we know that for all ε' sufficiently close to 1^-, f' has exactly two zeros t_1 and t_2. By Lemma 7, we know that $f'(t) > 0$ for $t \in (t_1, t_2)$. These imply that $f(t)$ is decreasing on $t \in (0, t_1)$, increasing on $t \in (t_1, t_2)$ and decreasing on $t \in (t_2, 1)$, which gives $\arg\max_{t\in[0,1]} f(t) \subseteq \{0, t_2\}$. Furthermore, since $f(0) = 0$ and $f'(t) < 0$ for $t \in (0, t_1)$, we know $f(t) < 0$ in $t \in (0, t_1)$. Also note that $f(1) = -1 < 0$. Therefore, depending on ε', the sign of $f(t)$ in $t \in (0,1)$ only has two possibilities: either $f(t) < 0$ for all $t \in (0,1)$ except possibly one point where $f(t) = 0$, or there exist τ_1 and τ_2 as described in (b). In the latter case we have $0 < t_1 < \tau_1 < t_2 < \tau_2 < 1$.

We now show the existence of such τ_1 and τ_2 for all ε' sufficiently close to 1^-. Since we have shown that $\arg\max_{t\in[0,1]} f(t) \subseteq \{0, t_2\}$ and $f(0) = 0$, it suffices to show $f(t_2) > 0$. Since $f'(t_2) = 0$, we have $f(t_2) > 0 \Leftrightarrow f(t_2) - t_2 \cdot f'(t_2) > 0 \Leftrightarrow [(1+\varepsilon')^3 - (1+\varepsilon')] t_2^{(1+\varepsilon')^2} > [(1-\varepsilon') - (1-\varepsilon')^3] t_2^{(1-\varepsilon')^2}$, which can be simplified into

$$\frac{(1+\varepsilon')^3 - (1+\varepsilon')}{(1-\varepsilon') - (1-\varepsilon')^3} > \frac{1}{t_2^{4\varepsilon'}} .$$

Since $\varepsilon' < 1$, it then suffices to show

$$1 + \frac{6}{\frac{2}{\varepsilon'} + \varepsilon' - 3} \geq \frac{1}{t_2^4} .$$

Observe that LHS $\to +\infty$ as $\varepsilon' \to 1^-$. It remains to show that t_2 is bounded away from 0 as $\varepsilon' \to 1^-$, i.e., $\liminf_{\varepsilon'\to 1^-} t_2(\varepsilon') > 0$. We claim that $\liminf_{\varepsilon'\to 1^-} t_2 \geq \frac{1}{2}$. To show this, we note that

$$\liminf_{\varepsilon'\to 1^-} f'(q, \varepsilon') = \liminf_{\varepsilon'\to 1^-} 1 - (1+\varepsilon')^3 \cdot q^{(1+\varepsilon')^2 - 1} - (1-\varepsilon')^3 \cdot q^{(1-\varepsilon')^2-1} = 1 - 2^3 \cdot q^3 ,$$

which equals zero when $q = \frac{1}{2}$.

The claim in (b) that $\tau_2(\varepsilon') \geq \frac{1}{3}$ follows directly from the above analysis since $t_2 < \tau_2$ and $\liminf_{\varepsilon'\to 1^-} t_2 \geq \frac{1}{2}$.

To show $\lim_{\varepsilon'\to 1^-} \tau_1(\varepsilon') = 0$, we claim that $\tau_1 \leq (1-\varepsilon')^{0.9}$ as $\varepsilon' \to 1^-$. Then it suffices to show that $f((1-\varepsilon')^{0.9}, \varepsilon') > 0$ for all $\varepsilon' \to 1^-$. We have

$$\frac{1}{(1-\varepsilon')^{0.9}} \cdot f((1-\varepsilon')^{0.9}, \varepsilon') = 1 - (1+\varepsilon')(1-\varepsilon')^{0.9[(1+\varepsilon')^2-1]} - (1-\varepsilon')^{1+0.9[(1-\varepsilon')^2-1]} ,$$

which tends to 1 as $\varepsilon' \to 1^-$. This implies (b).

We now show (c). First note that $f(0) = 0$ and $f(1) = -1$.

When $\varepsilon' = 1$, $f(t) = t - 2t^4$. In this case, we have $f(t) > 0$ for $t \in (0, 2^{-1/3})$ and $f(t) < 0$ for $t \in (2^{-1/3}, 1)$.

When $1 < \varepsilon' \leq 2$, by Lemma 7, we have $f'(t) = 1 + (\varepsilon'-1)^3 t^{(\varepsilon'-1)^2-1} - (\varepsilon'+1)^3 t^{(\varepsilon'+1)^2-1}$ being strictly decreasing on $t \in (0,1)$. Therefore the function $f(t)$ is concave. Since $\lim_{t\to 0^+} f'(t) > 0$, $f(0) = 0$ and $f(1) = -1 < 0$, the result follows by concavity.

When $2 < \varepsilon'$, again since $f(0) = 0$ and $f(1) = -1$, it suffices to show f is strictly increasing and then strictly decreasing on $t \in (0,1)$. Note that since $\lim_{t\to 0^+} f'(t) = 1 > 0$ and $\lim_{t\to 1^-} f'(t) < 0$, it then suffices to show $f'(t)$ is increasing and then decreasing on $(0,1)$. To show this, it suffices to show that if $f''(\hat{t}) = \frac{\partial}{\partial t} f(\hat{t}) < 0$ for some $\hat{t} \in (0,1)$, then $f''(t) < 0$ for all $t \in [\hat{t}, 1)$. Now, since

$$f''(\hat{t}) < 0 \Leftrightarrow \frac{(\varepsilon'-1)^3\left[(\varepsilon'-1)^2 - 1\right]}{(\varepsilon'+1)^3\left[(\varepsilon'+1)^2 - 1\right]} < \hat{t}^{(\varepsilon'+1)^2 - (\varepsilon'-1)^2} ,$$

and $\hat{t}^{(\varepsilon'+1)^2-(\varepsilon'-1)^2} < t^{(\varepsilon'+1)^2-(\varepsilon'-1)^2}$ for all $t \geq \hat{t}$, we conclude that $f''(t) < 0$ for all $t \in [\hat{t}, 1)$. So we are done.

□

Now we are in a position to prove Theorem 1.

Proof of Theorem 1. Let $t_j = e^{-v_j^2}$ for all $j \in [d]$. By Lemma 3 and Lemma 4, we have

$$\begin{aligned}\frac{dL_n}{dn} &= W \sum_{j\in[d]} \mu(j) \frac{\partial L(v_j, \varepsilon'_j)}{\partial v_j} \cdot \frac{dv_j}{dn} = \frac{2W}{\sqrt{\pi}} \sum_{j\in[d]} \mu(j) f(t_j, \varepsilon'_j) \cdot \frac{\mu(j)}{2\sqrt{2}\sigma(j)\sqrt{n}}\,, \\ &= \frac{W}{\sqrt{2n\pi}} \sum_{j\in[d]} \frac{\mu^2(j)}{\sigma(j)} f(t_j, \varepsilon'_j)\,.\end{aligned} \tag{6}$$

By part (a) of Lemma 5, when $\varepsilon < \delta_1 \min_{j\in[d]} \mu(j)$, we have for all $j \in [d]$, it holds that $\varepsilon'_j < \delta_1$ and thus $f(t_j, \varepsilon'_j) < 0$ for all $t \in (0, 1)$. Combining it with (6) yields $\frac{dL_n}{dn} < 0$.

When $\max_{j\in[d]} \mu(j) \leq \varepsilon$, we have for all $j \in [d]$, it holds that $1 < \varepsilon'_j$. It follows from part (c) of Lemma 5 that for all $j \in [d]$, there exists $\tau_2(\varepsilon'_j)$ such that $f(t_j, \varepsilon'_j) > 0 \;\forall\; t_j \in (0, \tau_2(\varepsilon'_j))$. Pick $\tau_2 = \min_j \tau_2(\varepsilon'_j)$. Then for all $j \in [d]$, we have $f(t_j, \varepsilon'_j) > 0$ when $t_j < \tau_2$. Since $t_j = e^{-v_j^2} = \exp(-\frac{n\mu^2(j)}{2\sigma^2(j)})$, when $\exp(-\frac{n\mu^2(j)}{2\sigma^2(j)}) < \tau_2$, or equivalently $n > 2\log\left(\frac{1}{\tau_2}\right) \max_{j\in[d]} \frac{\sigma^2(j)}{\mu^2(j)}$, we have $\frac{dL_n}{dn} > 0$.

When $\delta_2 \cdot \max_{j\in[d]} \mu(j) < \varepsilon < \min_{j\in[d]} \mu(j)$, we have for all $j \in [d]$, it holds that $\delta_2 < \varepsilon'_j < 1$. Then by part (b) of Lemma 5, for all $j \in [d]$, $\exists\; \tau_1(\varepsilon'_j)$ and $\tau_2(\varepsilon'_j)$ such that

$$f(t_j, \varepsilon'_j) \begin{cases} < 0 & \forall t \in (0, \tau_1(\varepsilon'_j))\,, \\ > 0 & \forall t \in (\tau_1(\varepsilon'_j), \tau_2(\varepsilon'_j))\,, \\ < 0 & \forall t \in (\tau_2(\varepsilon'_j), 1)\,, \end{cases} \tag{7}$$

where $\tau_1(\varepsilon'_j) \to 0^+$ as $\varepsilon'_j \to 1^-$ and $\tau_2(\varepsilon'_j) > \frac{1}{3}$, for all $j \in [d]$. Let $\tau_2 = \max_{j\in[d]} \tau_2(\varepsilon'_j) > \frac{1}{3}$, $\tau_1 = \min_{j\in[d]} \tau_1(\varepsilon'_j)$ and $\hat{\tau}_1 = \max_{j\in[d]} \tau_1(\varepsilon'_j)$. Note that since $\lim_{\varepsilon'_j \to 1^-} \tau_1(\varepsilon'_j) = 0$, without loss of generality we can assume $\hat{\tau}_1 < \frac{1}{3}$. It follows from (7) that for all $j \in [d]$

$$f(t_j, \varepsilon'_j) \begin{cases} < 0 & \forall t \in (0, \tau_1)\,, \\ > 0 & \forall t \in \left(\hat{\tau}_1, \frac{1}{3}\right)\,, \\ < 0 & \forall t \in (\tau_2, 1)\,. \end{cases} \tag{8}$$

Denote $\gamma = \frac{\mu(j)}{\sigma(j)}$ for all $j \in [d]$ since this ratio is fixed. Then we have $t_j = \exp\left(-\frac{\mu^2(j)n}{2\sigma^2(j)}\right) = \exp(-\gamma^2 n/2)$. Therefore we can choose $N_4 = \log(\tau_1^{-1}) \cdot \left(\frac{2}{\gamma^2}\right)$, $N_3 = \log(\hat{\tau}_1^{-1}) \cdot \left(\frac{2}{\gamma^2}\right)$, $N_2 = \log(3) \cdot \left(\frac{2}{\gamma^2}\right)$ and $N_1 = \log(\tau_2^{-1}) \cdot \left(\frac{2}{\gamma^2}\right)$ where $N_1 < N_2 < N_3 < N_4$ and the result follows from (6) and (8).

□

Proof of Corollary 2. From the proof of Theorem 1, in this simplified case we have $\tau_1 = \hat{\tau}_1$ and $\tau_2 = \tau_2(\varepsilon'_j)$ for all j. It follows that the thresholds $N_1,\ N_2,\ N_3$, and N_4 in Theorem 1 satisfy $N_1 = N_2$, and N_3 is no longer needed and can be replaced by N_4. Therefore only two thresholds are needed in Corollary 2. We denote the two thresholds as N_1 and N_2.

It remains to show $\lim_{\varepsilon \to \mu_0^-} N_2(\varepsilon) - N_1(\varepsilon) = +\infty$. From part (b) of Lemma 5 and (6), we know the derivative $\frac{dL_n}{dn}$ is positive when $t := \exp(-\frac{n\mu_0^2}{2\sigma_0^2}) \in (\tau_1, \tau_2)$, or equivalently $n \in \left(\log(\frac{1}{\tau_2})\frac{2\sigma_0^2}{\mu_0^2}, \log(\frac{1}{\tau_1})\frac{2\sigma_0^2}{\mu_0^2}\right)$. By (b) of Lemma 5, we know $\tau_1 \to 0^+$ as $\varepsilon \to \mu_0^-$ while τ_2 is bounded away from 0. This shows $\lim_{\varepsilon \to \mu_0^-} \log(\frac{1}{\tau_1}) - \log(\frac{1}{\tau_2}) = +\infty$ and completes the proof. □

2 PROOF OF LEMMA

In this section we give the proof of Lemma 3.

Let $f^* \in S_2$, i.e., f^* is a minimizer of $\sum_{i=1}^n \max_{||\tilde{x}_i - x_i||_\infty < \varepsilon} H\left(-y_i(\tilde{t}_i - f(\tilde{s}_i))\right)$ with the smallest ℓ_1 norm. To show S_2 is nonempty and such f^* does exist, we specify the form of f^*. We claim that f^* can take the following form

$$f^*(s) = \sum_{j=1}^N \alpha_j \mathbb{1}[s \in I_j], \tag{9}$$

where $I_j = (j - \varepsilon, j + \varepsilon)$, $j \in [N]$. Indeed, by definition of H, we know that the value of f^* outside those intervals I_j's won't change the value of $H\left(-y_i(\tilde{t}_i - f(\tilde{s}_i))\right)$. Therefore in order to attain the smallest possible ℓ_1 norm, we must have $f^*(s) = 0$ for all $s \notin \cup_j I_j$.

Note that by letting $\varepsilon < 1/2$, any two intervals have no overlap. To see why f^* is a constant function over each interval I_j, we consider three possible cases of the dataset $\{(x_i, y_i), i \in [n]\}$. For the first case, suppose that those data points with $s = j$ contain only positive points. Then in order to correctly classify these points with ε perturbation, we must have $f^*(s) \le \mu - \varepsilon$ for all $s \in I_j$. In order to minimize $||f^*||_1$, we would take $\alpha_j = \min\{0,\ \mu - \varepsilon\}$. Similarly, if those points purely consist of negative points, then $\alpha_j = \max\{0,\ -\mu + \varepsilon\}$. For the second case, suppose that those data points with $s = j$ contain both positive and negative points. Suppose the number of positive points exceeds the number of negative points. Then to correctly classify the positive points, we have $f^*(s) \le \mu - \varepsilon$ for all $s \in I_j$. To correctly classify the negative points, we have $f^*(s) \ge -\mu + \varepsilon$ for all $s \in I_j$. If $-\mu + \varepsilon \le 0 \le \mu - \varepsilon$, then $\alpha_j = 0$. Otherwise, if $-\mu + \varepsilon > \mu - \varepsilon$, then f^* can never simultaneously classify both classes correctly. It will choose to correctly classify the class with more points, which is the positive class. Then $\alpha_j = \mu - \varepsilon$. On the other hand, if negative class has more points, then $\alpha_j = -\mu + \varepsilon$. If the two class have equal number of points at $s = j$, then α_j can be either $-\mu + \varepsilon$ or $\mu - \varepsilon$. For the third case, assume no point in the training set has $s = j$. Then $\alpha_j = 0$.

We have now specified the form that $f^* \in S_2$ can take, which also indicates that S_2 is nonempty. We now show for all sufficiently small λ, $S(\lambda) = S_2$.

First we show $S(\lambda) \subseteq S_2$. Let $f \in S(\lambda)$. We want to show $f \in S$ and $||f||_1 \le ||\hat{f}||_1$ for all $\hat{f} \in S$. Suppose on the contrary that $f \notin S$. Then by definition of H, there exists $f^* \in S$ s.t.

$$\sum_{i=1}^n \max_{||\tilde{x}_i - x_i||_\infty < \varepsilon} H\left(-y_i(\tilde{t}_i - f^*(\tilde{s}_i))\right) \le \sum_{i=1}^n \max_{||\tilde{x}_i - x_i||_\infty < \varepsilon} H\left(-y_i(\tilde{t}_i - f(\tilde{s}_i))\right) - 1/2$$

and since S_2 is nonempty we can further assume f^* satisfies

$$||f^*||_1 \in \arg\min_{\hat{f} \in S} ||\hat{f}||_1.$$

Since $f \in S(\lambda)$, we then have $\lambda||f||_1 \le \lambda||f^*||_1 - 1/2$, which implies $||f^*||_1 \ge 1/2\lambda$. From above analysis we know f^* must take the form of Eq. (9) where $\alpha_j \le |\mu - \varepsilon|$, and I_j has length equal to 2ε. This implies $||f^*||_1 \le 2N\varepsilon|\mu - \varepsilon|$. Therefore, if we pick $\lambda < \frac{1}{4N\varepsilon|\mu - \varepsilon|}$, then such f^* cannot exist. Therefore, for all sufficiently small λ, we have $f \in S$.

Now we show $||f||_1 \le ||\hat{f}||_1$ for all $\hat{f} \in S$. Suppose on the contrary that there exists $f^* \in S$ such that $||f^*||_1 < ||f||_1$. However, since we have already shown

$$\sum_{i=1}^n \max_{||\tilde{x}_i - x_i||_\infty < \varepsilon} H\left(-y_i(\tilde{t}_i - f^*(\tilde{s}_i))\right) = \sum_{i=1}^n \max_{||\tilde{x}_i - x_i||_\infty < \varepsilon} H\left(-y_i(\tilde{t}_i - f(\tilde{s}_i))\right),$$

this would contradict the fact that $f \in S(\lambda)$. Therefore we have $S(\lambda) \subseteq S_2$.

To see $S_2 \subseteq S(\lambda)$ for all sufficiently small λ, we again pick $\lambda < \frac{1}{4N\varepsilon|\mu - \varepsilon|}$. Note that since $||f^*||_1 \le 2N\varepsilon|\mu - \varepsilon|$ for all $f^* \in S_2$, we have $\lambda||f^*||_1 < \frac{1}{2}$. Now suppose on the contrary that there exists $f \notin S_2$ such that

$$\sum_{i=1}^n \max_{||\tilde{x}_i - x_i||_\infty < \varepsilon} H\left(-y_i(\tilde{t}_i - f(\tilde{s}_i))\right) + \lambda||f||_1 < \sum_{i=1}^n \max_{||\tilde{x}_i - x_i||_\infty < \varepsilon} H\left(-y_i(\tilde{t}_i - f^*(\tilde{s}_i))\right) + \lambda||f^*||_1.$$

Since $f^* \in S_2$, we must have $\lambda||f||_1 < \lambda||f^*||_1 < \frac{1}{2}$. Now, if $\sum_{i=1}^n \max_{||\tilde{x}_i - x_i||_\infty < \varepsilon} H\left(-y_i(\tilde{t}_i - f(\tilde{s}_i))\right) \le \sum_{i=1}^n \max_{||\tilde{x}_i - x_i||_\infty < \varepsilon} H\left(-y_i(\tilde{t}_i - f^*(\tilde{s}_i))\right)$, this would contradict the fact that f^* is in $\arg\min_S ||f||_1$. Therefore we must have $\sum_{i=1}^n \max_{||\tilde{x}_i - x_i||_\infty < \varepsilon} H\left(-y_i(\tilde{t}_i - f(\tilde{s}_i))\right) > \sum_{i=1}^n \max_{||\tilde{x}_i - x_i||_\infty < \varepsilon} H\left(-y_i(\tilde{t}_i - f^*(\tilde{s}_i))\right)$. However, by definition of H, this implies

$$\begin{aligned}\sum_{i=1}^n \max_{||\tilde{x}_i - x_i||_\infty < \varepsilon} H\left(-y_i(\tilde{t}_i - f(\tilde{s}_i))\right) + \lambda ||f||_1 &\geq \sum_{i=1}^n \max_{||\tilde{x}_i - x_i||_\infty < \varepsilon} H\left(-y_i(\tilde{t}_i - f^*(\tilde{s}_i))\right) + \frac{1}{2} + \lambda ||f||_1 \\ &\geq \sum_{i=1}^n \max_{||\tilde{x}_i - x_i||_\infty < \varepsilon} H\left(-y_i(\tilde{t}_i - f^*(\tilde{s}_i))\right) + \lambda ||f^*||_1,\end{aligned}$$

which is a contradiction. Therefore $S_2 \subseteq S(\lambda)$. Altogether we have $S(\lambda) = S_2$.

3 PROOF OF THEOREM

In this section, we give the proof of Theorem 4.

The proof follows from the Lemma 3 and its proof. By Lemma 3, we have $S(\lambda) = S_2$ and we can consider the equivalent definition that $f_n^{\text{rob}} \in S_2$. From the proof of Lemma 3, we know f_n^{rob} must take the form of (9). Since $|\alpha_j| \leq |\mu - \varepsilon|$, when $\varepsilon < 2\mu$, we have $|\alpha_j| < \mu$ and thus $|f_n^{\text{rob}}(s)| < \mu$ for all $s \in \mathbb{R}$. For such f_n^{rob}, we have $H\left(-y\left(t - f_n^{\text{rob}}(s)\right)\right) = 0$ for all $(x, y) = (s, t, y)$ in the support of $\mathcal{D}_{2N}$. This implies $L_n = 0$ for all n.

Assume $2\mu < \varepsilon < 1/2$. Then $|\alpha_j|$ can take the value of either 0 or $|\mu - \varepsilon| > |\mu|$. When $\alpha_j = 0$, f_n^{rob} can classify both the positive and negative points at location $s = j$ correctly. When $|\alpha_j| > \mu$, then f_n^{rob} can only classify one of the two classes correctly. Note that $\alpha_j = 0$ if and only if there is no point with $s = j$ in the training set. Let the random variable $Z \in 0 \cup [N]$ denote the cardinality of the set $\{j \in [N] : \; s_i \neq j \text{ for all } i \in [n]\}$, which is a function of the training set $\{(x_i, y_i)\}_{i=1}^n$. Then the generalization error can be written as

$$L_n = \mathbb{E}_{\{(x_i,y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}_{2N}} \frac{N - Z}{N} = 1 - \frac{\mathbb{E}_{\{(x_i,y_i)\}_{i=1}^n} Z}{N}.$$

Note that $\mathbb{E}_{\{(x_i,y_i)\}_{i=1}^n} Z$ decreases as n increases. Therefore $L_n < L_{n+1}$ for all n.

4 FURTHER DETAILS ON GAUSSIAN MIXTURE WITH 0-1 LOSS

4.1 PROOF OF PROPOSITION

Here we give the proof of Proposition 5.

Proof of Proposition 5:. By (1), it suffices to show that under the 0-1 loss

$$\sum_{i=1}^n \max_{\tilde{x}_i \in B_{x_i}^\infty(\varepsilon)} \mathbb{1}[y_i(\tilde{x}_i - w) < 0] = \sum_{i=1}^n y_i \mathbb{1}[x_i' < w]\,. \tag{10}$$

Conditioning on whether there exists $\tilde{x}_i \in B_{x_i}^\infty(\varepsilon)$ such that $\mathbb{1}[y_i(\tilde{x}_i - w) < 0] = 1$ or not, one can deduce that

$$\arg\max_{\tilde{x}_i \in B_{x_i}^\infty(\varepsilon)} \mathbb{1}[y_i(\tilde{x}_i - w) < 0] \supseteq \arg\min_{\tilde{x}_i \in B_{x_i}^\infty(\varepsilon)} y_i(\tilde{x}_i - w) = \{x_i'\}\,,$$

and it follows that

$$\sum_{i=1}^n \max_{\tilde{x}_i \in B_{x_i}^\infty(\varepsilon)} \mathbb{1}[y_i(\tilde{x}_i - w) < 0] = \sum_{i=1}^n \mathbb{1}[y_i(x_i' - w) < 0] = \sum_{i=1}^n y_i \mathbb{1}[x_i' < w]\,.$$

□

4.2 TEST LOSS AND OPTIMAL TIEBREAK

To find the optimal tiebreaking in hingsight, we need to minimize the test loss over the model parameter w, which is given by Proposition 9.

Proposition 9. *The test loss of classifier w is given by*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_\mathcal{N}}[\mathbb{1}[y(x-w)<0]] = \frac{1}{2} + \frac{1}{2}\left(\Phi\left(\frac{w-\mu}{\sigma}\right) - \Phi\left(\frac{w+\mu}{\sigma}\right)\right), \quad (11)$$

where Φ is the CDF of the standard normal distribution. Furthermore, the minimizer of (11) *is $w=0$.*

Proposition 9 indicates that the optimal tiebreak in hindsight chooses the point closest to 0 (i.e., the point with the minimum absolute value) from (the closure of) the interval where w^* lies. This is because $w = 0$ minimizes the test loss in (11), and one can see that (11) increases as $|w|$ increases. Indeed, the derivative of (11) is given by $\frac{1}{2\sigma\sqrt{2\pi}}\left(\exp(-\frac{(w-\mu)^2}{2\sigma}) - \exp(-\frac{(w+\mu)^2}{2\sigma})\right)$, which is negative for $w < 0$ and positive for $w > 0$.

Proof of Proposition 9:. Conditioning on $y = \pm 1$, we have

$$\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mathcal{D}_\mathcal{N}}[\mathbb{1}[y(x-w)<0]] \\
&= \mathbb{P}(y=1)\cdot\mathbb{E}_{x|y=1}[\mathbb{1}[y(x-w)<0]] + \mathbb{P}(y=-1)\cdot\mathbb{E}_{x|y=-1}[\mathbb{1}[y(x-w)<0]] \\
&= \frac{1}{2}\cdot\mathbb{E}_{x\sim\mathcal{N}(\mu,\sigma)}[\mathbb{1}[x-w<0]] + \frac{1}{2}\cdot\mathbb{E}_{x\sim\mathcal{N}(-\mu,\sigma)}[\mathbb{1}[x-w>0]] \\
&= \frac{1}{2}\cdot\mathbb{P}_{z\in\mathcal{N}(0,1)}\left(z<\frac{w-\mu}{\sigma}\right) + \frac{1}{2}\cdot\mathbb{P}_{z\in\mathcal{N}(0,1)}\left(z>\frac{w+\mu}{\sigma}\right) \\
&= \frac{1}{2}\cdot\Phi\left(\frac{w-\mu}{\sigma}\right) + \frac{1}{2}\cdot\left[1-\Phi\left(\frac{w+\mu}{\sigma}\right)\right].
\end{aligned}$$

Since the derivative is $\frac{1}{2\sigma\sqrt{2\pi}}\left(\exp(-\frac{(w-\mu)^2}{2\sigma}) - \exp(-\frac{(w+\mu)^2}{2\sigma})\right)$, we see that $w^* = 0$ minimizes the above quantity. □