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

Yifei Min∗ Lin Chen† Amin Karbasi‡

Abstract

Despite remarkable success, deep neural networks are sensitive to human-imperceptible small perturbations on the data and could be adversarially misled to produce incorrect or even dangerous predictions. To circumvent these issues, practitioners introduced adversarial training to produce adversarially robust models whose predictions are robust to small perturbations to the data. It is widely believed that more training data will help adversarially robust models generalize better on the test data. In this paper, however, we challenge this conventional belief and show that more training data could hurt the generalization of adversarially robust models for the linear classification problem. We identify three regimes based on the strength of the adversary. In the weak adversary regime, more data improves the generalization of adversarially robust models. In the medium adversary regime, with more training data, the generalization loss exhibits a double descent curve. This implies that in this regime, there is an intermediate stage where more training data hurts their generalization. In the strong adversary regime, more data almost immediately causes the generalization error to increase.

1 Introduction

In recent years, deep neural networks have exhibited their superiority over traditional models in an abundance of machine learning tasks, e.g., image classification [Krizhevsky et al., 2012], speech recognition and language translation [Graves et al., 2013, Bahdanau et al., 2015], medical diagnosis [Lakhami and Sundaram, 2017, Xiao et al., 2019], text recognition and information extraction [Long et al., 2020, Mei et al., 2018, Wang et al., 2012], product recommendations [Zhang et al., 2020], online fraud detection [Pumsirirat and Yan, 2018], self-driving cars [Ramos et al., 2017], information security [Guan et al., 2019], among others. However, they can also be extremely vulnerable to adversarial, human-imperceptible data modifications [Szegedy et al., 2013, Carlini and Wagner, 2018, Kos et al., 2018]. This vulnerability is even more concerning and dangerous when neural networks are used in scenarios directly connected to human safety. For instance, medical diagnosis (misinterpreting medical images) or self-driving cars (misreading traffic signs). To circumvent these issues, practitioners introduce adversarial training in order to produce adversarially robust models [Huang et al., 2015, Shaham et al., 2018, Madry et al., 2017]. They are expected to be robust to minor data alteration and make consistent predictions.

There is a large body of work dedicated to adversarially robust models [Zhang and Zhu, 2019, Santurkar et al., 2019, Zhang et al., 2019, Bhagoji et al., 2019, Diochnos et al., 2019, Wei and Ma, 2019, Tsipras et al., 2019] proved the existence of a trade-off between the standard accuracy of a model and its robustness to adversarial perturbations. Ilyas et al. [2019] showed that adversarial

∗Department of Statistics and Data Science, Yale University. E-mail: yifei.min@yale.edu.
†Yale Institute for Network Science, Yale University. E-mail: lin.chen@yale.edu.
‡Yale Institute for Network Science, Yale University. E-mail: amin.karbasi@yale.edu.
examples could be directly attributed to non-robust features. Montasser et al. [2019] studied the learnability of an adversarially robust model. They established that any hypothesis class with finite VC dimension is robustly PAC learnable with an improper learning rule.

In this paper, we focus on the generalization performance of adversarially trained models on the standard (i.e., unperturbed) test data. Previous works studied the generalization of adversarially robust models from a variety of perspectives. Yin et al. [2019] and Kim and Loh [2018] explored the generalization of adversarially robust models from the perspective of Rademacher complexity which provides a bound on the generalization error. In contrast, our paper presents a precise characterization of the generalization performance, rather than a bound or an approximate result. Raghunathan et al. [2019] showed empirical evidence that adversarial training could hurt the standard accuracy, despite its improvement on the robust accuracy. In a recent paper, Chen et al. [2020] studied the influence of a larger training dataset upon the gap between the generalization performance of an adversarially robust model and a standard model. They found that more training data could result in expansion of the gap and denied the belief that more training data would help adversarially robust models reach a similar generalization performance to the standard model.

1.1 Our Contributions

In this paper, we investigate the generalization performance of adversarially trained models on standard test data. A conventional wisdom in machine learning is that a larger training dataset will result in better generalization on the test data. We provably establish a surprising and to some extent even paradoxical result that more training data can hurt the generalization of adversarially trained models. We consider a linear classification problem under two different loss functions: the linear loss and the 0-1 loss. For the linear loss, we identify three regimes of different adversary strengths, i.e., the weak, medium, and strong adversary regimes.

- In the strong adversary regime, the generalization of adversarially robust models deteriorates with more training data, except for a possible short initial stage where the generalization is improved with more data.

- The medium adversary regime is probably the most interesting one among the three regimes. In this regime, the evolution of the generalization performance of adversarially robust models could be a double descent curve. In particular, at the initial stage, the generalization loss on the test data is reduced with more training data. At the intermediate stage, however, the generalization loss increases as there is more training data (more data hurts the generalization of adversarial robust models). At the final stage, more training data improves the generalization performance.

- In the weak adversary regime, the generalization is consistently improved with more training data.

Similar weak and strong adversary regimes are observed for the 0-1 loss. The generalization error decreases in the weak adversary regime while decreases in the strong adversary regime. For medium adversary, there is first an increasing stage where the error increases as more training data is used. After this stage the error will keep going down.

2 Related Work

In this section, we briefly discuss some of the papers on the generalization of adversarially robust models and the double descent phenomenon, which are most relevant to our work.
Schmidt et al. [2018] show that adversarially robust models need more training data compared to their standard counterpart. They consider a Gaussian mixture model similar to ours and prove that the training of a robust model requires a training dataset with size $\Omega(d)$ where $d$ is the dimension of the data, whereas the standard model only need a constant number of data points. Bubeck et al. [2019] study a binary classification problem under a statistical query setting and show that to train a robust classifier one need exponentially (in dimension $d$) many queries, while only polynomially many to train a standard one. As discussed in Section 1, the main difference between theirs and ours is that we quantify the training dynamic in terms of size of the training dataset. Cullina et al. [2018] and Diochnos et al. [2019] consider a PAC-learning setting. Cullina et al. [2018] provide a polynomial (in the VC dimension) upper bound for the sample complexity, while Diochnos et al. [2019] gives an lower bound for the sample complexity which is exponential in the dimension of the input.

The strength of the adversary is crucial in adversarial training. Theoretically, Dohmatob [2019] show that a classifier with high standard accuracy can inevitably be fooled by a strong adversary. Empirically, Papernot et al. [2016] and Tsipras et al. [2019] find that a strong adversary can lead to deteriorated standard accuracy for robust models. Ilyas et al. [2019] find that the model tends to learn non-robust features and omit robust ones during the adversarial training if the adversary is too strong.

The double descent phenomenon has been studied by several authors. Belkin et al. [2019a,b] and Mei and Montanari [2019] have proved the existence of double descent curves for the generalization error under certain model setting. However, we would like to remark that the double descent curve they consider is in terms of the number of parameters (model complexity), while ours is sample-wise. Empirically, Nakkiran et al. [2019] also discover a sample-wise double descent phenomenon.

## 3 Main Results

Throughout this paper, let $[n]$ be a shorthand notation for $\{1, 2, \ldots, n\}$. We consider a binary classification problem where each data point $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$ is generated by a Gaussian mixture model. More specifically, the distribution for the data generation is specified by $y \sim \text{Unif}(\{\pm 1\})$ and $x \mid y \sim \mathcal{N}(y\mu, \Sigma)$, where $\mu(j) \geq 0$ for all $j \in [d]$ and $\Sigma = \text{diag}(\sigma^2(1), \sigma^2(2), \ldots, \sigma^2(d))$. In the remaining parts we denote this distribution by $(x, y) \sim D_N$.

For this data generation procedure, we analyze two different loss functions. One is the linear loss function $\ell(x, y; w) = -y\langle w, x \rangle$ where $w \in \Theta = \{w \in \mathbb{R}^d \mid \|w\|_\infty \leq W\}$ and $W > 0$ is a positive constant, similar to [Chen et al., 2020] Yin et al., 2019, Khim and Loh, 2018. In this setting, we provably show that there are three possible regimes (weak, medium and strong adversary regimes), in which more training data can help, double descend, or hurt generalization of the adversarially trained model, respectively. The other loss function that we study is the 0-1 loss function $\ell(x, y; w) = \mathbb{1}[y(x - w) < 0]$, where $w \in \mathbb{R}$ and data is one-dimensional. For this loss function, we empirically show that the generalization error can also increase with the size of the training dataset.

With either loss function, the robust classifier is defined as follows:

$$w_n^{\text{rob}} = \arg\min_{w \in \Theta} \sum_{i=1}^{n} \max_{\tilde{x} \in B^\infty(\epsilon)} \ell(\tilde{x}, y_i; w),$$

where $\Theta$ is the parameter space where $w$ lives and $B^\infty(\epsilon) \triangleq \{\tilde{x} \in \mathbb{R}^d \mid \|\tilde{x} - x\|_\infty \leq \epsilon\}$ is an $\ell^\infty$ ball centered at $x$ and with radius $\epsilon$. The radius $\epsilon$ characterizes the strength of the adversary. A larger $\epsilon$ means a stronger adversary. In the case of linear loss, we set $\Theta = \{w \in \mathbb{R}^d \mid \|w\|_\infty \leq W\}$, where
Figure 1: This figure illustrates the three regimes (i.e., weak, medium, and strong adversary regimes) and the results of Theorem 1 in these regimes. In the weak adversary regime, more training data always improves generalization. The medium adversary regime exhibits a double descent curve. If we denote the size of the training dataset by $n$, when $n \leq N_1$ (the initial stage), more training data improves the generalization; when $N_2 < n < N_3$ (the intermediate stage), generalization is hurt by more data; when $n \geq N_4$, more data helps with generalization again. In the strong adversary stage, when the size of training size is sufficiently large ($n > N_5$), generalization deteriorates with more training data.

$W > 0$ is a positive constant. In the case of 0-1 loss, we set $\Theta = \mathbb{R}$. This robust classifier minimizes the robust loss, or equivalently, maximizes the robust reward (i.e., negative loss).

The generalization error of the robust classifier under the linear loss is given by

$$L_n = \mathbb{E}_{\{(x, y)_i\}_{i=1}^n \sim \mathcal{D}_N} \left[ \mathbb{E}_{(x, y) \sim \mathcal{D}_N} [\ell(x, y; w_{n}^{\text{rob}})] \right],$$

where the inner expectation is over the randomness of the test data point and the outer expectation is over the randomness of the training dataset. The test and training data are assumed to be independently sampled from the same distribution. The generalization error can be interpreted as the expected loss of the robust model over standard test data.

### 3.1 Linear Loss

In this subsection, we consider the linear loss $\ell(x, y; w) = -y \langle w, x \rangle$ where $w \in \Theta = \{w \in \mathbb{R}^d \|w\|_\infty \leq W\}$ and $W$ is a positive constant. In this setting, we have the robust classifier

$$w_{n}^{\text{rob}} = \arg \min_{\|w\|_\infty \leq W} \sum_{i=1}^{n} \max_{\tilde{x}_i \in B_{2\varepsilon}^\infty} (-y_i \langle w, \tilde{x}_i \rangle) = \arg \max_{\|w\|_\infty \leq W} \sum_{i=1}^{n} \min_{\tilde{x}_i \in B_{2\varepsilon}^\infty} y_i \langle w, \tilde{x}_i \rangle.$$  

We study how the generalization error of the robust model evolves as the size of the training dataset changes, i.e., the dependence of $L_n$ on $n$. The generalization error of the robust classifier under linear loss is given by

$$L_n = \mathbb{E}_{\{(x, y)_i\}_{i=1}^n \sim \mathcal{D}_N} \left[ \mathbb{E}_{(x, y) \sim \mathcal{D}_N} [-y \langle w_{n}^{\text{rob}}, x \rangle] \right].$$

For the Gaussian classification problem under the linear loss, we identify that the behavior of $L_n$ exhibits a phase transition which is determined by the strength of the adversary. Our main result is summarized by Theorem 1.
Theorem 1 (Proof in Section 5.1). Given $n$ i.i.d. training data points $(x_i, y_i) \sim D_N$, if the robust classifier is defined by (3) and its generalization error is defined by (4), then there exist $0 < \delta_1 < \delta_2 < 1$, such that

(a) If $0 < \varepsilon < \delta_1 \cdot \min_{j \in [d]} \mu(j)$, then $L_n < L_{n-1}$ for all $n$. That is, the loss $L_n$ monotonically decreases as the number of training points $n$ increases.

(b) If $\delta_2 \cdot \max_{j \in [d]} \mu(j) < \varepsilon < \min_{j \in [d]} \mu(j)$, and we further assume that $\frac{\mu(j)}{\sigma(j)}$ is the same for all $j$, then there exist $N_1 < N_2 < N_3 < N_4$ such that

$$L_n \begin{cases} < L_{n-1} & \text{for } 0 < n \leq N_1, \\ > L_{n-1} & \text{for } N_2 < n < N_3, \\ < L_{n-1} & \text{for } N_4 \leq n. \end{cases}$$

(c) If $\max_{j \in [d]} \mu(j) \leq \varepsilon$, then there exists $N_5$ such that $L_n > L_{n-1}$ for all $n > N_5$.

Theorem 1 verifies the existence of three possible regimes during the commonly used adversarial training procedure and gives conditions for when the phase transition between these regimes will take place. Part (a) identifies the weak regime, showing that when the strength of the adversary $\varepsilon$ is small compared to the signal $\mu$, the generalization error decreases as the size of the training dataset increases. In this regime, the generalization benefits from the use of a large training set. This regime is illustrated by the first plot in Fig. 1. We see that curve is always decreasing.

However, as the adversary becomes stronger, we reach the medium regime and things change. Part (b) proves the existence of a double descent curve for the generalization error. It shows that when $\varepsilon$ becomes larger and approaching the signal in magnitude, the generalization error will first decrease as as more training data is used. But surprisingly once it reaches a certain point, it will start increasing as we feed more data. This increasing stage continues until the dataset size reaches some threshold $N_2$ and then the error will decrease again. The medium adversary regime is illustrated by the second plot in Fig. 1, where the three stages are clearly marked by three different colored areas.

If the adversary’s strength reaches the same level as the signal or becomes even stronger, then for all sufficiently large $n$, the generalization error monotonically decreases as the number of training points increases. This strong regime is described in part (c) of the theorem and illustrated by the third plot in Fig. 1. Note that despite the decreasing stage near the very beginning, the loss keeps going up after the threshold $N_5$.

Furthermore, we see that in the medium regime, the length of the increasing stage is given by $N_3 - N_2$, according to part (b) of Theorem 1. We would like to remark that the model can have an arbitrarily long increasing stage, which depends on the adversary’s strength. To better interpret this idea and the meaning behind Theorem 1, we consider the following special case where $\mu(j) = \mu_0$ and $\sigma(j) = \sigma_0$ for all $j \in [d]$. In this special case, it can be shown that in the medium regime, as $\varepsilon$ approaches the signal strength $\mu_0$, the increasing stage grows and can be arbitrarily long.

Corollary 2 (Proof in Section 5.1). Under the same assumption as Theorem 1 and further assuming that $\mu(j) = \mu_0$ and $\sigma(j) = \sigma_0$ for all $j \in [d]$, we have

(a) If $0 < \varepsilon < \delta_1 \mu_0$, then $L_n < L_{n-1}$ for all $n$.

(b) If $\delta_2 \mu_0 < \varepsilon < \mu_0$, then there exist $N_1(\varepsilon) < N_2(\varepsilon)$ such that

$$L_n \begin{cases} < L_{n-1} & \text{for } 0 < n \leq N_1, \\ > L_{n-1} & \text{for } N_1 < n < N_2, \\ < L_{n-1} & \text{for } N_2 \leq n, \end{cases}$$
and \( \lim_{\varepsilon \to \mu_0^-} N_2(\varepsilon) - N_1(\varepsilon) = +\infty. \)

(c) If \( \mu_0 \leq \varepsilon \), then there exists \( N_3(\varepsilon) \) such that \( L_n > L_{n-1} \) for all \( n > N_3 \).

Part (a) and (c) of Corollary 2 are a re-statement of corresponding parts of Theorem 1 in the simplified setting. Part (b) additionally states that as \( \varepsilon \) increases towards \( \mu_0 \), the length of the increasing stage goes to infinity. In this setting, the three regimes are clearly marked by the thresholds \( \delta_1 \mu_0, \delta_2 \mu_0 \) and \( \mu_0 \).

Figure 2: The test loss versus the size of the training dataset under the linear loss and the one-dimensional \((d = 1)\) Gaussian data generation model described in Section 3.1. The parameters of the Gaussian data model are set as follows: \( \mu_0 = 1 \) and \( \sigma_0 = 2 \). In each plot, the solid curves correspond to robust models and the dashed curve corresponds to the standard model.

Fig. 2 illustrates the behavior of the generalization error in this simplified setting. In the simulation we set the parameters as \( d = 1, \mu_0 = 1 \) and \( \sigma_0 = 2 \) (for all three plots). Fig. 2a shows the weak adversary regime. We see that the generalization error maintains a decreasing trend when \( \varepsilon \) is as large as half the signal strength. In Fig. 2b it is clear that the generalization error has a double descent curve. At first there is a decreasing stage, which is followed by an increasing stage. Also observe that as \( \varepsilon \) becomes larger, the error increases faster during the increasing stage. The error will finally start decreasing as the size of training dataset reaches the second decreasing stage. On the contrary, in the strong adversary regime the increasing stage lasts forever and the error keeps increasing no matter how much data is provided, as illustrated by Fig. 2c.

### 3.2 0-1 Loss

We consider the 0-1 loss for the one-dimensional linear classification problem (i.e., \( d = 1 \)). Let us recall the data generation process. To generate a data point \((x, y) \in \mathbb{R} \times \{\pm 1\} \), first, we sample \( y \) from \( \text{Unif}(\{\pm 1\}) \). Second, given \( y \in \{\pm 1\} \), we sample \( x \sim \mathcal{N}(y\mu, \sigma) \). The classifier is represented by a single real number \( w \in \mathbb{R} \). The classifier predicts \( y = +1 \) if \( x \geq w \) and predicts \(-1\) otherwise.

Given a data point \((x, y)\), the 0-1 loss of classifier \( w \) is given by \( \ell(x, y; w) = 1[y(x - w) < 0] \).

If the training dataset is \( \{(x_i, y_i)\}_{i=1}^n \), we define the neutralized dataset \( \{(x'_i, y_i)\}_{i=1}^n \) that satisfies \( x'_i = x_i - y_i\varepsilon \) for \( \forall i \in [n] \). In other words, for a positive sample \((x_i, y_i = 1)\), we obtain its neutralized sample by shifting \( x_i \) to the negative direction by \( \varepsilon \), i.e., \( x'_i = x_i - \varepsilon \); for a negative sample \((x_i, y_i = -1)\), its neutralized sample is obtained by shifting \( x_i \) to the positive direction by \( \varepsilon \), i.e., \( x'_i = x_i + \varepsilon \). We see that the dataset remains unchanged after neutralization if \( \varepsilon = 0 \). With this definition, the robust classifier can be expressed as the following.

**Proposition 3** (Proof in Section 5.2). *Given the training dataset \( \{(x_i, y_i)\}_{i=1}^n \) and the neutralized dataset \( \{(x'_i, y_i)\}_{i=1}^n \), the robust classifier can be expressed as the following.*
dataset \( \{(x'_i, y_i)\}_{i=1}^n \), the robust classifier is given by

\[
w_{rob} \in \arg \min_{w \in \mathbb{R}^n} \sum_{i=1}^n y_i \mathbbm{1}[x'_i < w]. \tag{5}
\]

There is a tiebreaking issue in light of Proposition 3. To see this, let \( s \) be the permutation of \([n]\) such that \( x'_{s(1)} \leq x'_{s(2)} \leq \cdots \leq x'_{s(n)} \). The \( n \) points divide the real line into \( n+1 \) intervals: \((-\infty, x'_{s(1)}], (x'_{s(i)}, x'_{s(i+1)}] \) for \( 1 \leq i \leq n-1 \), and \((x'_{s(n)}, \infty)\). Let \( w^* \) be a minimizer of (5). If \( w^* \) lies in any of the above \( n+1 \) intervals, then any other point on the same interval is also a minimizer, since at these two points the objective function has the same value. Therefore, a tiebreaking procedure is required here.

We study the following two tiebreaking methods: agnostic tiebreak and the optimal tiebreak in hindsight. If \( w^* \in (x'_{s(i)}, x'_{s(i+1)}) \), the agnostic tiebreak chooses \( w_{rob} \) uniformly at random from the interval. If \( w^* > x'_{s(n)} \), it chooses \( w_{rob} \) arbitrarily close to \( x'_{s(n)} \) from above. If \( w^* \leq x'_{s(1)} \), it chooses \( w_{rob} = x'_{s(1)} \). To find the optimal tiebreaking in hindsight, we need to minimize the test loss over the model parameter \( w \), which is given by Proposition 4.

**Proposition 4** (Proof in Section 5.3). The test loss of classifier \( w \) is given by

\[
\mathbb{E}_{(x,y)\sim D_X}[\mathbbm{1}[y(x - w) < 0]] = \frac{1}{2} + \frac{1}{2} \left( \Phi \left( \frac{w - \mu}{\sigma} \right) - \Phi \left( \frac{w + \mu}{\sigma} \right) \right), \tag{6}
\]

where \( \Phi \) is the CDF of the standard normal distribution. Furthermore, the minimizer of (6) is \( w = 0 \).

Proposition 4 indicates that the optimal tiebreak in hindsight chooses the point closest to 0 (i.e., the point with the minimum absolute value) from the interval where \( w^* \) lies. This is because we know that \( w = 0 \) minimizes the test loss in (6).

![Figure 3: The test loss vs. the size of the training dataset in the 0-1 loss setting. The solid curves correspond to robust models and the dashed curve corresponds to the standard model.](image)

(a) Agnostic tiebreak

(b) Optimal tiebreak in hindsight

Fig. 3 illustrates the test loss versus the size of the training dataset under both the agnostic tiebreak and the optimal tiebreak in hindsight. In Fig. 3 we set \( \mu = \sigma = 1 \) and use the same set of values for \( \epsilon \) for both tiebreaking methods. We have three observations. First, the generalization error is increasing in \( n \) when \( \epsilon \) is larger than the signal strength. This confirms the existence of the strong adversary regime under the 0-1 loss. Second, for small enough \( \epsilon \) (e.g. \( \epsilon \leq 0.5 \)), the
generalization error is decreasing in \( n \) (more precisely after \( n = 3 \)). For the medium adversary where \( \varepsilon \) is in between 0.7 and 1.0, the curve has an increasing stage followed by a decreasing stage, which is very similar to what we see in Fig. 2b.

4 Conclusion

In this paper we prove that the generalization of adversarially trained models does not always benefit from a larger training dataset. When the adversary is relatively weak, the model can still properly learn via adversarial training and the generalization error keeps decreasing as more data is used. When the adversary is stronger and the perturbation becomes comparable to the signal, the generalization error shows a double descent behavior. When the perturbation exceeds the level of the signal, the model cannot effectively learn from data, which results in an increasing generalization error as the number of training points grows. One possible explanation for this result is that if the adversary is very strong then it can perturb the data so much that it totally erases the signal of the original distribution. Therefore, when adversarially-trained models try to account for this different distribution, they can no longer generalize to the unperturbed test data that is drawn from the original distribution.

The goal of adversarial training is to produce robust models that provide protection against attacks that make perturbations to the data at test time. While protection against such attacks is undoubtedly important, we still want our robust models to perform well on unperturbed data. However, our results indicate that there are scenarios in which it is impossible for current approaches to achieve low generalization error. In fact, if the adversary is strong enough, the generalization error for current adversarially robust models will become worse and worse as we add more data. This is in direct contradiction to one of the primary tenets of machine learning, which is that more data should help us learn better. Our findings suggest that the current adversarial training framework may not be ideal and that fundamentally new ideas may be required to develop models that can reliably perform well on both perturbed and unperturbed test sets.

5 Proofs

5.1 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., 2020] that gives the closed form solution for the robust classifier.

Proposition 5 (Lemma 10 in [Chen et al., 2020]). 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 \( w_{\text{rob}} = W \text{sign}(u - \varepsilon \text{sign}(u)) \), where \( u = \frac{1}{n} \sum_{i=1}^{n} y_i x_i \).

First, we define the error function [Andrews, 1998] \( \text{erf}(\cdot) : \mathbb{R} \to \mathbb{R} \) by

\[
\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} \, dt, 
\]

and it has the following property.

Lemma 6. If \( z \sim \mathcal{N}(0, 1) \), we have

\[
P(z < x) = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{x}{\sqrt{2}} \right) \right].
\]
Proof of Lemma 7. In light of the density of the standard normal distribution and by a change of variable, we have
\[ P(z < x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} \, dt = \frac{1}{2} + \frac{\sqrt{2}}{\sqrt{2\pi}} \int_0^{x/\sqrt{2}} e^{-s^2} \, ds = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{x}{\sqrt{2}} \right) \right]. \]

In addition, we define the function \( L(\cdot, \cdot) : \mathbb{R}^2 \rightarrow \mathbb{R} \) by
\[ L(v, \varepsilon') = \text{erf}(v) + \text{erf}(v \varepsilon' - 1) - \text{erf}(v \varepsilon' + 1). \] (8)

For all \( j \in [d] \), we define
\[ v_j = \frac{\sqrt{n} \mu(j)}{\sqrt{2} \sigma(j)}, \quad \varepsilon'_j = \frac{\varepsilon}{\mu(j)}, \] (9)
where \( \mu(j) \) and \( \sigma(j) \) are defined in the data generation process described at the beginning of Section 3.

Lemma 7 gives the expression for the generalization error.

Lemma 7. Suppose that the generalization error is defined as in (4). Then we have
\[ L_n = W \sum_{j \in [d]} \mu(j) L(v_j, \varepsilon'_j), \]
where \( v_j \) and \( \varepsilon'_j \) are defined in (9).

Proof of Lemma 7. By (4), Proposition 5 and the independence between test and training data, we have
\[ L_n = -E_{\{(x_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{D}_N} \left[ E_{(x, y) \sim \mathcal{D}_N} \left[ y \langle w_n^{rob}, x \rangle \right] \right] = -E_{\{(x_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{D}_N} \left[ \langle w_n^{rob}, \mu \rangle \right] \]
\[ = -W \cdot \sum_{j \in [d]} \mu(j) E_{\{(x_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{D}_N} \left[ \text{sign} (u(j) - \varepsilon \text{sign} (u(j))) \right] \]
Since \( y_i x_i \sim \mathcal{N}(\mu, \Sigma) \), we have \( u \sim \mathcal{N}(\mu, \Sigma) \), and it follows that
\[ L_n = -W \cdot \sum_{j \in [d]} \mu(j) E_{u(j) \sim \mathcal{N}(\mu(j), \sigma^2(j)/n)} \left[ \text{sign} (u(j) - \varepsilon \text{sign} (u(j))) \right]. \]

Denote \( I_j = -E_{u(j) \sim \mathcal{N}(\mu(j), \sigma^2(j)/n)} \left[ \text{sign} (u(j) - \varepsilon \text{sign} (u(j))) \right] \). Then we have
\[ I_j = P(u(j) < -\varepsilon) - P(-\varepsilon < u(j) < 0) + P(0 < u(j) < \varepsilon) - P(\varepsilon < u(j)) \]
\[ = 1 - 2P(-\varepsilon < u(j) < 0) - 2P(\varepsilon < u(j)) \]
\[ = 1 - 2P \left( \frac{\varepsilon - \mu(j)}{\sigma(j)} \sqrt{n} < z < \frac{-\mu(j)}{\sigma(j)} \sqrt{n} \right) - 2P \left( \frac{\varepsilon - \mu(j)}{\sigma(j)} \sqrt{n} < z \right) \]
\[ = 1 - 2 \left[ P \left( z < \frac{\varepsilon + \mu(j)}{\sigma(j)} \sqrt{n} \right) - P \left( z < \frac{\mu(j)}{\sigma(j)} \sqrt{n} \right) \right] - 2 \left[ 1 - P \left( z < \frac{\varepsilon - \mu(j)}{\sigma(j)} \sqrt{n} \right) \right], \]
where \( z \) is a standard normal random variable. By Lemma 6 we have
\[ I_j = \text{erf} \left( \frac{\mu(j)}{\sqrt{2} \sigma(j)} \sqrt{n} \right) + \text{erf} \left( \frac{\varepsilon - \mu(j)}{\sqrt{2} \sigma(j)} \sqrt{n} \right) - \text{erf} \left( \frac{\varepsilon + \mu(j)}{\sqrt{2} \sigma(j)} \sqrt{n} \right) \]
\[ = \text{erf}(v_j) + \text{erf}(v_j \varepsilon'_j - 1) - \text{erf}(v_j \varepsilon'_j + 1) = L(v_j, \varepsilon'_j), \]
which implies that $L_n = W \sum_{j \in [d]} \mu(j) L(v_j, \varepsilon'_j)$. \hfill \square

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 7 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

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

In Lemma 8 we compute the partial derivative of $L$.

**Lemma 8.** Let $t = e^{-v^2}$ and $f$ be defined as in (5.1). 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 8.** By (7) we have

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

and it follows by (8) that

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

The proof of Theorem 1 follows from studying the derivative $\frac{dL_n}{dn}$. Lemma 8 implies that the derivative depends on the sign of the function $f$. We investigate the sign of $f$ in Lemma 9.

**Lemma 9.** There exist $0 < \delta_1 \leq \delta_2 < 1$ 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 $\varepsilon'$ 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

$$\lim_{\varepsilon' \to 1^-} \tau_1(\varepsilon') = 0,$$

$$\tau_2(\varepsilon') \geq \frac{1}{3}.$$

(c) When $1 \leq \varepsilon'$, $f(t, \varepsilon')$, 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 9 uses the following Lemma 10 and Lemma 11. To make it concise, whenever we fix $\varepsilon'$ in the context, we omit $\varepsilon'$ and write $f(t) = f(t, \varepsilon')$ and $f'(t) = f'(t, \varepsilon').$

**Lemma 10.** 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

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

The proof of Lemma 10 follows from direct computation. Using Lemma 10 we obtain Lemma 11.

**Lemma 11.** 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 $1 \leq \varepsilon' < 2$, $f'(t)$ is strictly decreasing for $t \in (0, 1)$.

**Proof of Lemma 11.** 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

$$t_0 = \left[ \frac{(1 + \varepsilon')^3}{1 - \varepsilon'} \left( \frac{2 + \varepsilon'}{2 - \varepsilon'} \right) \right]^\frac{1}{3} \quad (10).$$

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

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

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

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

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

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

\hfill \Box

A direct application of Lemma 11 gives the following Lemma 12.
Lemma 12. For all $0 < \varepsilon' < 1$ sufficiently close to 1, $f'(t)$ has exactly two zeros on $t \in (0,1)$.

Proof of Lemma 12. By Lemma 11 we know that $f'(t)$ is strictly increasing on $t \in (0, t_0)$ and strictly decreasing on $(t_0, 1)$. Recall that Lemma 10 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 $\varepsilon'$ sufficiently close to 1. We define

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

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

$$f'(t_0) = 1 - (1 + \varepsilon')^3 A^{-\frac{1}{2} + \frac{\varepsilon'}{4}} - (1 - \varepsilon')^3 A^\frac{1}{2} - \frac{\varepsilon'}{4}.$$  \hspace{1cm} (11)

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

$$\lim_{\varepsilon' \to 1^-} (1 - \varepsilon')^3 A^\frac{1}{2} - \frac{\varepsilon'}{4} = \lim_{\varepsilon' \to 1^-} (1 - \varepsilon')^3 \left(1 + \frac{2 \varepsilon'}{2 - \varepsilon'}\right)\left[1 - \frac{\varepsilon'}{4\varepsilon'}\right] = 0.$$  \hspace{1cm} (12)

Therefore we conclude that $f'(t_0) > 0$ as $\varepsilon' \to 1^-$. We denote the two zeros in Lemma 12 by $t_1 = t_1(\varepsilon')$ and $t_2 = t_2(\varepsilon')$ where $t_1 < t_2$.

Now we are ready to prove Lemma 9.

Proof of Lemma 9. We show (a) first. Note that for any fixed $\varepsilon' < 1$, $f(0) = 0$. Therefore it suffices to show that for any $\varepsilon'$ sufficiently close to 0, the derivative $f'(t) < 0$. Since by Lemma 11 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 $\varepsilon'$ sufficiently close to 0.

In light of (10), $f'(t_0) < 0$ is equivalent to

$$1 - (1 + \varepsilon')^3 \left[\left(\frac{1 + \varepsilon'}{1 - \varepsilon'}\right)^3 \left(\frac{2 + \varepsilon'}{2 - \varepsilon'}\right)\right] - (1 - \varepsilon')^3 \left[\left(\frac{1 + \varepsilon'}{1 - \varepsilon'}\right)^3 \left(\frac{2 + \varepsilon'}{2 - \varepsilon'}\right)\right] < 0.$$  \hspace{1cm} (13)

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

$$\frac{2 + \varepsilon'}{2 - \varepsilon'} < \frac{(1 - \varepsilon')^3}{1 - (1 - \varepsilon')^3}.$$  \hspace{1cm} (14)

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 12 we know that for all $\varepsilon'$ sufficiently close to $1^-$, $f'$ has exactly two zeros $t_1$ and $t_2$. By Lemma 11 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 \( \varepsilon' \), 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 \( \tau_1 \) and \( \tau_2 \) 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 \( \tau_1 \) and \( \tau_2 \) for all \( \varepsilon' \) sufficiently close to 1\(^-\). Since we have shown that \( \arg \max_{t \in [0, 1]} f(t) \leq \{0, t_2\} \) and \( f(0) = 0 \), it suffices to show \( f(t_2) > 0 \). Since \( f'(t_2) = 0 \), we have \( f(t_2) > 0 \iff f(t_2) - t_2 \cdot f'(t_2) > 0 \iff [(1 + \varepsilon')^3 - (1 + \varepsilon')]t_2^{1+\varepsilon'} > [(1 - \varepsilon') - (1 - \varepsilon')^3]t_2^{1-\varepsilon'^2} \), 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

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

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., \( \lim \inf_{\varepsilon' \to 1^-} t_2(\varepsilon') > 0 \). We claim that \( \lim \inf_{\varepsilon' \to 1^-} t_2 \geq \frac{1}{2} \). To show this, we note that

\[
\lim \inf_{\varepsilon' \to 1^-} f'(q, \varepsilon') = \lim \inf_{\varepsilon' \to 1^-} 1 - (1 + \varepsilon')^3 \cdot q^{(1+\varepsilon')^2} - (1 - \varepsilon')^3 \cdot q^{(1-\varepsilon')^2} - 1 = 2^3 \cdot q^3,
\]

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 \( \lim \inf_{\varepsilon' \to 1^-} t_2 \geq \frac{1}{2} \).

To show \( \lim \inf_{\varepsilon' \to 1^-} \tau_1(\varepsilon') = 0 \), 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

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

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 \([1] \) we have \( f(t) = 1 + (\varepsilon' - 1)^3(\varepsilon' - 1)^2 - 1 - (\varepsilon' + 1)^3(\varepsilon' + 1)^2 - 1 \) 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 \( f''(\hat{t}) = \frac{9}{2}\hat{t} f(\hat{t}) < 0 \) for some \( \hat{t} \in (0, 1) \), then \( f''(t) < 0 \) for all \( t \in [\hat{t}, 1) \). Now, since

\[
f''(\hat{t}) < 0 \iff \frac{(\varepsilon' - 1)^3 [(\varepsilon' - 1)^2 - 1]}{(\varepsilon' + 1)^3 [(\varepsilon' + 1)^2 - 1]} < \frac{9}{2}(\varepsilon' + 1)^2 - (\varepsilon' - 1)^2,
\]

and \( \frac{9}{2}(\varepsilon' + 1)^2 - (\varepsilon' - 1)^2 < \frac{9}{2}(\varepsilon' + 1)^2 - (\varepsilon' - 1)^2 \) for all \( t \geq \hat{t} \), we conclude that \( f''(t) < 0 \) for all \( t \in [\hat{t}, 1) \). So we are done.

\[\square\]

Now we are in a position to prove Theorem \([1] \).

Proof of Theorem 4. Let \( t_j = e^{-v_j^2} \) for all \( j \in [d] \). By Lemma 7 and Lemma 8, we have

\[
\frac{dL_n}{dn} = W \sum_{j \in [d]} \mu(j) \frac{\partial L(v_j, \tau_j)}{\partial v_j} \cdot \frac{dv_j}{dn} = \frac{2W}{\sqrt{\pi}} \sum_{j \in [d]} \mu(j) f(t_j, \tau_j) \cdot \frac{\mu(j)}{2\sqrt{2\sigma(j)}\sqrt{n}},
\]

(12)

By part (a) of Lemma 9, when \( \varepsilon < \delta_1 \min_{j \in [d]} \mu(j) \), we have for all \( j \in [d] \), it holds that \( \tau_j^* < \delta_1 \) and thus \( f(t_j, \tau_j^*) < 0 \) for all \( t \in (0, 1) \). Combining it with (12) 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 \( 1 < \tau_j^* \). It follows from part (c) of Lemma 9 that for all \( j \in [d] \), there exists \( \tau_j^* \) such that \( f(t_j, \tau_j^*) > 0 \) \( \forall t \in (0, \tau_j^*) \). Pick \( \tau_2 = \min_j \tau_2(\tau_j^*) \). Then for all \( j \in [d] \), we have \( f(t_j, \tau_j^*) > 0 \) when \( t_j < \tau_2 \). Since \( t_j = e^{-v_j^2} = \exp(-\frac{n\mu^2(j)}{2\sigma(j)}) \), when \(\exp(-\frac{n\mu^2(j)}{2\sigma(j)}) < \tau_2 \), 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 \( \delta_2 < \tau_j^* < 1 \). Then by part (b) of Lemma 9 for all \( j \in [d] \), \( \exists \tau_1(\tau_j^*) \) and \( \tau_2(\tau_j^*) \) such that

\[
f(t_j, \tau_j^*) = \begin{cases} < 0 & \forall t \in (0, \tau_1(\tau_j^*)), \\ > 0 & \forall t \in (\tau_1(\tau_j^*), \tau_2(\tau_j^*)), \\ < 0 & \forall t \in (\tau_2(\tau_j^*), 1), \end{cases}
\]

(13)

where \( \tau_1(\tau_j^*) \rightarrow 0^+ \) as \( \tau_j^* \rightarrow 1^\rightarrow \) and \( \tau_2(\tau_j^*) > \frac{1}{3} \), for all \( j \in [d] \). Let \( \tau_2 = \max_{j \in [d]} \tau_2(\tau_j^*) > \frac{1}{3} \), \( \tau_1 = \min_{j \in [d]} \tau_1(\tau_j^*) \) and \( \tilde{\tau}_1 = \max_{j \in [d]} \tau_1(\tau_j^*) \). Note that since \( \lim_{\tau_j^* \rightarrow 1^-} \tau_1(\tau_j^*) = 0 \), without loss of generality we can assume \( \tilde{\tau}_1 < \frac{1}{3} \). It follows from (13) that for all \( j \in [d] \)

\[
f(t_j, \tau_j^*) = \begin{cases} < 0 & \forall t \in (0, \tilde{\tau}_1), \\ > 0 & \forall t \in (\tilde{\tau}_1, \frac{1}{3}), \\ < 0 & \forall t \in (\tau_2, 1), \end{cases}
\]

(14)

Denote \( \gamma = \frac{\mu(j)}{\sigma(j)} \) for all \( j \in [d] \) since this ratio is fixed. Then we have \( t_j = \exp(-\frac{n\mu^2(j)}{2\sigma(j)}) = \exp(-\gamma^2 n/2) \). Therefore we can choose \( N_4 = \log(\tau_1^{-1}) \cdot \left( \frac{2}{\tau_1^2} \right) \), \( N_3 = \log(\tilde{\tau}_1^{-1}) \cdot \left( \frac{2}{\tilde{\tau}_1^2} \right) \), \( N_2 = \log(3) \cdot \left( \frac{2}{\tau_2^2} \right) \) and \( N_1 = \log(\tau_2^{-1}) \cdot \left( \frac{2}{\tau_2^2} \right) \) where \( N_1 < N_2 < N_3 < N_4 \) and the result follows from (12) and (14).

Proof of Corollary 2. From the proof of Theorem 4 in this simplified case we have \( \tau_1 = \tilde{\tau}_1 \) and \( \tau_2 = \tau_2(\tau_j^*) \) for all \( j \). It follows that the thresholds \( \tau_1, N_2, N_3, \) 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 9 and (12), we know the derivative \( \frac{dL_n}{dn} \) is positive when \( t := \exp(-\frac{n\mu^2}{2\sigma^2}) \in (\tau_1, \tau_2) \), or equivalently \( n \in \left( \log(\frac{1}{\tau_1}) \frac{2\sigma^2}{\mu^2}, \log(\frac{1}{\tau_2}) \frac{2\sigma^2}{\mu^2} \right) \). By (b) of Lemma 9 we know \( \tau_1 \to 0^+ \) as \( \varepsilon \to \mu_0^- \) while \( \tau_2 \) is bounded away from 0. This shows \( \lim_{\varepsilon \to \mu_0^-} \frac{1}{\log(\frac{1}{\tau_1})} - \frac{1}{\log(\frac{1}{\tau_2})} = +\infty \) and completes the proof.
5.2 Proof of Proposition 3

By (1), it suffices to show that under the 0-1 loss

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

(15)

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

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

and it follows that

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

5.3 Proof of Proposition 4

Conditioning on \( y = \pm 1 \), we have

\[
\begin{align*}
E_{(x, y) \sim D_N} \left[ \mathbb{1}[y(x - w) < 0] \right] & = P(y = 1) \cdot E_{x|y=1} \left[ \mathbb{1}[y(x - w) < 0] \right] + P(y = -1) \cdot E_{x|y=-1} \left[ \mathbb{1}[y(x - w) < 0] \right] \\
& = \frac{1}{2} \cdot E_{x \sim N(\mu, \sigma)} [\mathbb{1}[x - w < 0]] + \frac{1}{2} \cdot E_{x \sim N(-\mu, \sigma)} [\mathbb{1}[x - w > 0]] \\
& = \frac{1}{2} \cdot P_{z \sim N(0, 1)} \left( z < \frac{w - \mu}{\sigma} \right) + \frac{1}{2} \cdot P_{z \sim N(0, 1)} \left( z > \frac{w + \mu}{\sigma} \right) \\
& = \frac{1}{2} \cdot \Phi \left( \frac{w - \mu}{\sigma} \right) + \frac{1}{2} \cdot \left[ 1 - \Phi \left( \frac{w + \mu}{\sigma} \right) \right].
\end{align*}
\]

We see that \( w^* = 0 \) minimizes the above quantity.

Acknowledgements

We would like to thank Marko Mitrovic for his help in preparation of this paper.

References

Larry C Andrews. *Special functions of mathematics for engineers*, volume 49. Spie Press, 1998.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

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.
Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems*, pages 7496–7508, 2019.

Sebastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, pages 831–840, 2019.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE, 2018.

Lin Chen, Yifei Min, Mingrui Zhang, and Amin Karbasi. More data can expand the generalization gap between adversarially robust and standard models. *arXiv preprint arXiv:2002.04725*, 2020.

Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. Pac-learning in the presence of adversaries. In *Advances in Neural Information Processing Systems*, pages 230–241, 2018.

Dimitrios I Diochnos, Saeed Mahloujifar, and Mohammad Mahmoody. Lower bounds for adversarially robust pac learning. *arXiv preprint arXiv:1906.05815*, 2019.

Elvis Dohmatob. Generalized no free lunch theorem for adversarial robustness. In *International Conference on Machine Learning*, pages 1646–1654, 2019.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

Yushuo Guan, Yuanxing Zhang, Lin Chen, and Kaigui Bian. A neural attack model for cracking passwords in adversarial environments. In *2019 IEEE/CIC International Conference on Communications in China (ICCC)*, pages 183–188. IEEE, 2019.

Ruitong Huang, Bing Xu, Dale Schuurmans, and Csaba Szepesvári. Learning with a strong adversary. *arXiv preprint arXiv:1511.03034*, 2015.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

Justin Khim and Po-Ling Loh. Adversarial risk bounds for binary classification via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 36–42. IEEE, 2018.

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

Paras Lakhani and Baskaran Sundaram. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2):574–582, 2017.
Shangbang Long, Yushuo Guan, Kaigui Bian, and Cong Yao. A new perspective for flexible feature gathering in scene text recognition via character anchor pooling. *arXiv preprint arXiv:2002.03509*, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Hongyuan Mei, Sheng Zhang, Kevin Duh, and Benjamin Van Durme. Halo: Learning semantics-aware representations for cross-lingual information extraction. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 142–147, 2018.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. *arXiv preprint arXiv:1902.04217*, 2019.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2019.

Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*, 2016.

Apapan Pumsirirat and Liu Yan. Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine. *International Journal of advanced computer science and applications*, 9(1):18–25, 2018.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. In *ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena*, 2019.

Sebastian Ramos, Stefan Gehrig, Peter Pinggera, Uwe Franke, and Carsten Rother. Detecting unexpected obstacles for self-driving cars: Fusing deep learning and geometric modeling. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1025–1032. IEEE, 2017.

Shibani Santurkar, Andrew Ilyas, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Image synthesis with a single (robust) classifier. In *Advances in Neural Information Processing Systems*, pages 1260–1271, 2019.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pages 5014–5026, 2018.

Uri Shaham, Yutaro Yamada, and Sahand Negahban. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 307:195–204, 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.
Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In ICLR, 2019.

Tao Wang, David J Wu, Adam Coates, and Andrew Y Ng. End-to-end text recognition with convolutional neural networks. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pages 3304–3308. IEEE, 2012.

Colin Wei and Tengyu Ma. Improved sample complexities for deep networks and robust classification via an all-layer margin. arXiv preprint arXiv:1910.04284, 2019.

Jun Xiao, Yuanxing Zhang, Kaigui Bian, Guopeng Zhou, and Wei Yan. Denxfpn: Pulmonary pathologies detection based on dense feature pyramid networks. In ICASSP, pages 1234–1238. IEEE, 2019.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In ICML, pages 7085–7094, 2019.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In International Conference on Machine Learning, pages 7472–7482, 2019.

Tianyuan Zhang and Zhanxing Zhu. Interpreting adversarially trained convolutional neural networks. In International Conference on Machine Learning, pages 7502–7511, 2019.

Yuanxing Zhang, Pengyu Zhao, Yushuo Guan, Lin Chen, Kaigui Bian, Lingyang Song, Bin Cui, and Xiaoming Li. Preference-aware mask for session-based recommendation with bidirectional transformer. In ICASSP, 2020.